

Table of Contents

Section I. Introduction and General Essays	6
I.1 Approach to the Evidence Based Medicine Rotation (Ray Klein, Jonathan Ross)	6
I.2 Introductory Essentials- (Rebecca Wood).....	10
I.3 Cultural Competence in Evidence-Based Practice: Enhancing Synergy (Sirey H. Zhang, GSM4)....	13
I.4 Clinical Decision Making, Gut Feeling or Hard Rules? (Jacob Markwood, GSM4).....	18
Section II. Types of Studies	21
II.1 A Primer on the Design of Studies – (Jacqueline Raicek).....	21
II.2 Factorial Design, Main Effect, and Interactions (Yi Zhang)	22
II.3 Randomized Controlled Trial (RCT) Design: Key Elements in the Gold Standard of EBM (Keegan O’Hern).....	27
II.4 RCTs: Strengths and Limitations (Mariah Evarts).....	32
II.5 Systematic Reviews and Meta-Analyses (Alex Donovan)	34
II.6 Cluster Randomized Controlled Trials: What are they, when can they be used, and what biases might they introduce? (Emmalynn Moore, GSM4).....	37
II.7 Network Meta-Analysis- Explanation and Interpretation of a Unique Tool for EBM (David Styren).....	41
II.8 Designing a study- comparing superiority and non-inferiority studies (Charlie Calliff, GSM4) ..	50
II.9 Non-Inferiority Trials (Lukas Emery)	54
II.10 Pragmatic Clinical Trials- What Are They? (Priya Katari)	59
II.11 Assessing Pragmatism of Clinical Trials (Diana Lee, GSM4).....	60
II.12 Phases of New Drug Investigation Trials– (Katie Kozacka, GSM4).....	61
II.13 Understanding Endpoints with an emphasis on cancer trials (David Lakomy).....	64
Section III. Fundamental Research Methods and Statistics.....	70
III.1 The Bell Curve – What is a “normal distribution” and why does it matter? (Chad Y. Lewis, GSM4) 70	
III.2 The normal distribution and data analysis- Ben Seifer	73
III.3 Understanding Odds Ratios and Relative Risk Ratios (Barry Howe).....	84
III.4 Using Odds Ratio vs. a Hazard Ratio vs Relative Risk? (Erica Wadas).....	86
III.5 Statistical Bias (Anthony Bambara).....	88

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

III.6	A Cartoon Introduction to Type I and Type II Error (Adam Eddington, GSM 4).....	90
III.7	Statistical Error (Brenton Nash, Lizzy Schink, comments by Ken Phelps)	99
III.8	Statistical Significance- not as simple as $p < 0.05$ (Luke Mayer, GSM4)	108
III.9	Estimating Sample Size (Haley Moulton, GSM4)	111
III.10	The Rationale Behind Choosing the Appropriate Sample Size in Randomized Controlled Trials (Bill Rayburn).....	113
III.11	Blinding in Randomized Controlled Trials (Julia Harrison, GSM4)	116
III.12	Randomization (Chris Del Prete).....	118
III.13	Intention to Treat vs. As-Treated Analysis (Jen Frampton and Resham Ramkissoon)	122
III.14	Primary vs Secondary Outcomes, Dichotomizing, and Selection of Endpoints (Kayla Hatchell, GSM4)	123
III.15	Clinical Outcome Assessments and nuances of trial design in Neurology (Dennis Obat, GSM4)	126
III.16	Understanding the Continuing Conundrum of Continuous Variables (Swathi Krishnan, GSM4)	133
III.17	Confounding and how to mitigate its impact (Ashley Dunkle, GSM4)	135
III.18	Confounding and Effect Modification- why you need to know (Fili Bogdanic)	142
III.19	Coming to Terms with Composite Measures (Mallory Perez, GSM4)	147
III.20	Evaluation of Screening Tests (Alex Fiorentino)	155
III.21	Cancer Screening: Elements of an effective screening tool and other considerations- (Eric Jayne, GSM4)	158
III.22	Challenges of Diagnostic Testing and Risk for Bias (Rebecca Robbins)	164
III.23	Bayes Theorem (Malachy Sullivan and Alex Briand).....	169
III.24	Beyond Bayes: Some Issues in Diagnostic Reasoning; or, Towards an Evidence-based Framework for Diagnostic Reasoning (Stephen Conn GSM4)	174
III.25	The Placebo Effect- should we pay attention? (Devin van Dyke, GSM4)	179
III.26	The Big Data Paradox: A Conundrum of Abundance and Accuracy (Maria Malik GSM4)	182
Section IV.	Advanced Research Methods and Statistics	187
IV.1	Receiver Operating Curve Basics (Karim Farrag and Julia Lake)	187
IV.2	Kaplan-Meier Curves- Comparing Event/Survival Between Experimental and Control Groups (Taeha Kim and Rachel Griffith)	194
IV.3	The Cox Proportional Hazards Model (Art Kehas)	202
IV.4	Aggregation of Data in Meta-Analyses and How to Assess for Robustness of the Results (Chelsea Gaviola, GSM4).....	203
IV.5	Heterogeneity (Richie Huynh).....	206

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

IV.6	Forest Plots (Richie Huynh).....	208
IV.7	Regression analysis – a. Introduction to Linear Regression Analysis (Jiyong Lee).....	211
IV.8	Regression Analysis- b. Understanding and Using Logistic Regression (Muhammad Khan GSM4) 214	
IV.9	Introduction to Propensity Score Matching (Jiazuo Henry Feng).....	219
IV.10	What is Chi-square (X^2) testing? (Lauren Bernal, GSM4).....	222
IV.11	One-Way Analysis of Variance (ANOVA)—what is it, when is it used, and sample calculation (Marie Syku, GSM4)	228
IV.12	Post Hoc Analysis - What, Why, How, and What to Worry About? (Linda Morris, GSM4) ..	236
IV.13	Bonferroni Correction: What is it and when to use it? (Ahmed El Hussein, GSM3).....	240
IV.14	Inter-Rater Reliability and the Kappa Statistic (Vidal Villela).....	244
IV.15	Interim Analyses: When is it justified to prematurely terminate a Clinical Trial? (Navjot Sobti) 248	
IV.16	Sensitivity Analysis in Clinical Trials- (Meghan Freed, GSM4)	251
IV.17	Dealing with Missing Data- what’s a person to do? (Daniel Forsman, GSM4)	254
IV.18	Multiple Imputation and Controlled Multiple Imputation: Examples from the OPTION-DM trial (Will Carroll, GSM4)	259
IV.19	Reporting and Interpreting Economic Analysis (Ashley Baronner).....	263
IV.20	The Internists Guide to Choosing the Correct Statistical Test (J.D. Nuschke III).....	267
Section V.	Finding, Appraising, and Applying Evidence	273
V.1	Health literacy and numeracy – an essential feature of evidence based medicine- Caroline Lombardo.....	273
V.2	Obtaining High-Quality Studies and Clinical Guidelines: A User-Friendly Overview (Alexander Kettering, GSM4).....	275
V.3	Levels of Evidence and Recommendations- USPSTF and AAFP- (Aditya Kulkarni)	280
V.4	Search Strategies-From PICO to Primary Literature (Amogh Karnik)	285
V.5	Troubleshooting your search for evidence (Gwen Caffrey).....	291
V.6	An Algorithm to Assess Study Quality (Bianca Di Cocco, GSM4)	293
V.7	Critical Appraisal for Randomized Controlled Trial (Joan Chandra).....	296
V.8	Comparing and Contrasting Two or More Studies (Susan Wang)	298
V.9	Assessing the Risk of Bias of Randomized Controlled Trials in Systematic Reviews and Meta-Analyses (Chris Lindholm).....	300
V.10	Systematic Reviews and Assessing the Quality of Evidence with GRADE (Briana Goddard, GSM4) 302	
V.11	Questioning Quality of Qualitative Research (Sarah Baranes, GSM4).....	305

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

V.12	Measures of Impact (John Hon, GSM4)	311
V.13	History, Ethics, and Current State of Pediatric Research (Hira Haq, GSM4).....	313
V.14	Evidence Based Medicine in Pediatrics: Unique Challenges and Tools to Overcome Them – (Sarah Banerji, GSM4)	316
V.15	Applying Population-based Studies to the Individual Patient (Fatima Haidar, GSM 4).....	318
V.16	Statistical Incorporation of Patient Preferences and Values (Matt Wesley)	324
V.17	Minimal Clinically Important Difference – (Meg Hanley, GSM4).....	326
V.18	Decision Aids and Shared Decision Making (George S. Wang and Jesus Mendez Jr, GSM4) ...	332
V.19	Health Literacy and Numeracy: Tactics to Improve Communication and Patient Understanding of EBM (Lily Greene, GSM4).....	342
V.20	Why is it so difficult to prove mortality as an endpoint? The challenge of studying mortality in the critical care setting (Ashley Baronner)	347
V.21	How to assess treatment efficacy in solid tumor - an introduction to RECIST criteria – (Yuanzhen Cao)	349
V.22	Race and Ethnicity in EBM and Biomedical Research (Maya DeGroot, GSM4)	353
V.23	The Translational Highway- narrowing the gap between research and practice (Shantum Misra)	359
V.24	A new therapy gains FDA approval, then what? (Jon Pirruccello).....	363
V.25	Emergency Use Approval: What, How and Why it is used (Angela Lee, GSM4).....	365
V.26	Integrating Evidence-Based Medicine into Journal Club (Simrun Bal)	368
Section VI.	Integrating Diverse Sources of Information.....	375
VI.1	A Dive into Diabetes Management (Patrick Puliti)	375
VI.2	Utilizing Evidence-Based Medicine for Rare Diseases (Kyla Rodgers, GSM4).....	388
VI.3	Understanding Falls in the Elderly Utilizing EBM (Xingyi Li, GSM4).....	396
VI.4	Evidence Based Medicine and Pregnancy (Janae McGuirk, GSM4).....	406
VI.5	My experience with EBM and SGLT2 Inhibitors (Thomas Palladino, GSM 4)	408
VI.4	Evidence Based Medicine and Substance Use Disorders (Nikki Ratnapala, GMS4)	418
VI.5	EBM in Surgical Trials- Identifying unique challenges (Anna Witkin, GSM4).....	425
VI.6	Evidence-Based Psychiatry- History, Benefits, and Harms of the DSM-V (Rachel Brown, GSM4)	429
VI.7	The Blinded Leading the Blind – Unique Methodologic Challenges in Psychiatric Research and the Implications for Modern Psychedelic Research (Joseph Tella, GSM4).....	435
VI.8	Conducting Research with Native American Communities: Barriers and Considerations (Chenin Ryan, GSM4)	441

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Section I. Introduction and General Essays

In the beginning....

I.1 Approach to the Evidence Based Medicine Rotation (Ray Klein, Jonathan Ross)

The evidence-based medicine (EBM) rotation is a unique opportunity to hone and solidify a set of skills that will remain invaluable throughout a clinical career. During this rotation, residents develop skills to form clinical questions, find the strongest available evidence, critically appraise the relevant research, interpret study findings, and summarize the evidence in a way that helps readers make clinical decisions. There is no single “right” approach to the EBM rotation, but the following may provide a useful framework.

1. Remind Yourself of the Basics

Spend the initial portion of the rotation reading *Evidence-Based Medicine: How to practice and teach it* by Straus et al. The book is a short and easy read that will provide you with a solid foundation for the rest of the rotation. This is a useful refresher that covers everything from tips on forming an answerable clinical question to a review of essential statistics. Also take some time to review this EBM guide, which is filled with useful information created by residents who have previously participated in the EBM rotation.

2. Form a new PICO Question or Rebuild the PICO Question from a Research Article

There are many paths that may lead you to a research article. If you are interested in a particular subspecialty, you might be aware of some key journal articles within the field that you’ve been meaning to review. A patient may ask a question prompting you to search the literature for an answer. A new issue of the *New England Journal of Medicine* (or any other major medical journal) may have a research article you find interesting. However you find a journal article, remember the importance of creating a PICO question. If you are trying to answer a new clinical question, form your PICO question before you even begin to search through PubMed. If you are summarizing an interesting article you’ve already found, re-create the authors’ PICO question before you begin the summary by identifying each of the following:

Patient / Problem-P

Intervention-I

Comparison-C

Outcome-O

While it can be tempting to bypass the creation of a PICO question, establishing this framework at the beginning of the process will payback major dividends. Understanding the PICO framework will make it easier to find a relevant article, and it will also simplify the summary process.

3. Complete Any Necessary Background Reading

If you aren't already familiar with the topic of a journal article, it is often necessary to do some background reading. For example, if your article describes the effects of PCSK-9 inhibitors, you may need to quickly relearn their mechanism of action. Furthermore, you may need to quickly review the findings of previous research on PCSK-9 inhibitors. As always, UpToDate is an excellent resource for this type of background reading.

4. Interpret and Summarize the Research Article in the EBM Database

The goal is to create a succinct review that will help readers understand the fundamental question the article answers, the magnitude of the findings, the quality of the study, and the generalizability of the findings to a specific patient. Consider including the following in your EBM database summaries.

Question:

Restate the primary question answered by the research article you've found. Be sure to include important buzz words in this box, as anything included here is searchable within the EBM database.

Patients:

Describe the patient population, the number in the control arm and the intervention arm, inclusion criteria, exclusion criteria, and any important baseline demographics of the patients in the study. Include a description of the intervention and the control. List the outcomes that will appear below. This information will help readers determine how generalizable the results may be to their particular patient. For example, if 97% of the patients in a study are Caucasian, a reader would need to think critically about whether or not the results can be generalized to a patient of a different race. After describing the patient population, briefly describe the study protocol: How were the patients randomized? What happened to the intervention group? What happened to the control group? The duration of the study? The per cent follow-up?

Quality:

Describe the high-quality and low-quality characteristics of the study. Important factors to consider include randomization, blinding, sample size, length of follow up, intention-to-treat analysis, funding source, methodology flaws, etc. If not obvious, describe how a certain study characteristic may eliminate or create a source of bias.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Description of Intervention:

This should be a *one-line* description of what happened to the intervention group. For example, if half of the participants received liraglutide 1.8mg SubQ daily and the other half received placebo subQ injections, write “Liraglutide 1.8mg SubQ daily” in the intervention box. If necessary, a longer/more detailed description of the intervention should be included in the “Patients” box above.

Description of Control:

A *one-line* description of what was administered to the control group, e.g., placebo, or the comparison drug.

Outcomes:

Statistically describe any major outcomes from the study and any important adverse effects. If there are more than two study arms in your article, you will need to choose the two most relative to compare in this summary. Everything should be reported in terms of EER (experimental event rate) and CER (control event rate), which can then be used to calculate RRR and NNT (number needed to treat, or NNH, number needed to harm). You will need to know the number of patients in each arm, and then use the EBM calculator to find the confidence intervals for RRR and NNT. Quickly glancing at the number needed to treat/harm allows anyone reading your summary to get a quick sense of the magnitude of the study findings, particularly because you will have calculated the 95% CI (confidence interval). Remember, the EBM calculator reports the RRR and the CIs in a way that requires you to multiply by 100 before entering into the appropriate fields in the EBM database (one does not need to do the same for the NNT- use the numbers derived as is).

Randomized Controlled Trial Calculator

	Outcome		No Outcome	
Experimental	<input type="text" value="100"/>	A	<input type="text" value="650"/>	B
Control	<input type="text" value="200"/>	C	<input type="text" value="550"/>	D

Results

Chi-squared	40.838	p-value:	0
	Estimate	95% CI	
RRR	0.5	[0.378 to 0.598]	
ARR	0.133	[0.093 to 0.173]	
NNT	8	[11 to 6]	

So, in this example, the EER is 13.3%, the CER is 26.7%, and the RRR is 50% (but enter 50 into the data box in the New Study field because % is already embedded) and its confidence interval is 37.8 to 59.8

The ARR also needs to be multiplied by 100, thus the ARR is 13.3 with a CI of 9.3 to 17.3

1/ARR is the NNT which in the example is 7+ (rounded to 8 here) with a confidence interval of 11 to 6.

Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine

New Study field for Outcome

Outcome	EER	CER	RRR	CI	NNT	CI	
Death	13.3%	26.7%	50.2%	(.8 to 5)	7	(11 to 6)	<input type="checkbox"/> Harm <input type="checkbox"/> NS

(Remember- if your calculations use an outcome in which the result is better (higher) such as survival rather than death, the RRR is really a RRI, and you will need to multiply the RRR in the EBM Calculator line by -100 in order to make it transferable to the EBM database- that includes the CIs as well.)

Significance:

The first step in understanding the significance of the results is to establish the essential background. Start by *briefly* describing any essential pathophysiology and previous research in the field. For example, if you are summarizing an article on a new type of immuno-modulating chemotherapy, it would be nice to briefly remind the reader how the drug works. Furthermore, if your study is a follow up to previous research on the same drug, quickly note that in this section.

After briefly establishing the essential background, interpret the significance of the study outcomes. This is an opportunity to evaluate the importance and quality of the research. Is this a ground-breaking and practice-changing study? How big (or small) is the effect size? Are the findings generalizable to the relevant patient population? Are there major limitations that should temper enthusiasm? Is there another upcoming study on the topic we should watch out for in the next few years? This final section is the place to concisely describe the “take home message” of a journal article.

As you complete the evidence-based medicine rotation, you will no doubt have developed your own approach, but hopefully this proves to be a useful starting ground. At the end of the rotation, you should feel considerably more fluent and facile regarding the practice of EBM, be able to communicate more effectively with colleagues and patients alike, and hopefully establish some habits that will be helpful in promoting life-long learning. What follows are chapters written by residents who have taken this elective as a way to consolidate their learning and to contribute to your own learning. Enjoy!

I.2 Introductory Essentials- (Rebecca Wood)

Definitions:

Sensitivity: The probability of a disease person testing positive. Tests with a high sensitivity are used for screening as they may yield false positive results but do not miss people with the disease (low false negative rate).

Specificity: The probability of a non-diseased person testing negative. Tests with high specificity are used to confirm a disease is present.

Positive predictive value (PPV): If the test is positive, what is the probability that the patient has the disease? Depends on prior probability (or pre-test probability) and sensitivity/specificity of the test. The higher the prior probability, the greater the PPV. An overly sensitive test yields more false positive results and has a lower PPV.

Negative Predictive Value (NPV): If the test is negative, what is the probability that the patient does not have the disease? A high NPV is very important for a screening test. Also depends on prior probability and sensitivity/specificity. The more sensitive the test, the fewer number of false negative results and the higher the NPV.

Sensitivity and Specificity*

Result of Test Investigated	Result of Gold Standard Test	
	Disease Positive	Disease Negative
Positive (+)	TP (a)	FP (b)
Negative (-)	FN (c)	TN(d)

TP= True positive
 FP= False positive
 TN= True negative
 FN= False negative

Sensitivity= $a/a + c$ or $TP/TP+FN$ ↓

Specificity= $d/b + d$ or $TN/FP+TN$ ↑

Positive Predictive Value $PPV=a/a+b$ or $TP/TP+FP$ →

Negative Predictive Value $NPV= d/d+c$ or $TN/TN+FN$ ←

Likelihood Ratio Positive= sensitivity/1- specificity**

Likelihood Ratio Negative= 1- sensitivity/specificity**

*Sensitivity and Specificity are characteristics of the test, and do not vary with changes in prevalence or with changes in pre-test probability.

**Likelihood ratio is a way of combining the test characteristics (sensitivity and specificity) into a single measure.

When combined with odds, the LR (likelihood ratio) generates the post-test odds:

$$\text{Pre-test odds} \times \text{LR} = \text{Post-test odds}$$

And converting probability to odds is

$$\text{Odds} = \text{Probability} / 1 - \text{Probability}$$

$$\text{Probability} = \text{Odds} / 1 + \text{Odds}$$

Sensitivity helps rule OUT (SNOUT) Specificity helps rule IN (SPIN)		
Parameter	Definition	Calculation
Sensitivity	The probability of a diseased person testing positive	$\frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$
Specificity	The probability of a non-diseased person testing negative	$\frac{\text{True negatives}}{\text{True negatives} + \text{False positives}}$
Positive Predictive Value	The probability that disease is present given a positive result	$\frac{\text{True positives}}{\text{True positives} + \text{False positives}}$
Negative Predictive Value	The probability that disease is absent given a negative result	$\frac{\text{True negatives}}{\text{True negatives} + \text{False negatives}}$
Positive likelihood ratio	A ratio representing the likelihood of having the disease given a positive result	$\frac{\text{Sensitivity}}{1 - \text{Specificity}}$
Negative likelihood Ratio	A ratio representing the likelihood of having a disease given a negative result	$\frac{1 - \text{Sensitivity}}{\text{Specificity}}$

Other terms you will come across:

EER=Experimental event rate: outcome present/total in group exposed to experimental agent

CER=Control event rate: outcome present/total in group not exposed to experimental agent

ARR=Absolute risk reduction: EER-CER [Can also have ARI or ABI (absolute benefit increase)]

RRR=EER-CER divided by CER [can also have RRI or RBI (relative benefit increase)]

NNT= Number needed to treat: $1/\text{ARR}$ [Can also have NNH (number needed to harm) or NNS (number needed to screen)]

Relative Risk

RCTs or Prospective Cohort		Case outcomes	Control outcomes
Exposure	Yes	a	b
	No	c	d

Relative Risk = $\frac{a/(a+b)}{c/(c+d)}$ or $\frac{\text{exposed outcomes yes/all exposed}}{\text{not exposed yes/all not exposed}}$

Notes:

I.3 Cultural Competence in Evidence-Based Practice: Enhancing Synergy (Sirey H. Zhang, GSM4)

In a bustling clinic, Dr. Emily Lawson, an endocrinologist, struggled to help Mr. Patel, a man of Indian descent who was vegetarian, and who had diabetes. Despite following evidence-based dietary advice, Mr. Patel's blood sugar remained uncontrolled.

His traditional Indian diet differed significantly from the Western-based recommendations, making it hard for him to adhere to the prescribed low-carb, lean-protein plan.

Mr. Patel, feeling frustrated and culturally disconnected from his food, approached Dr. Lawson about this discordance, but she initially resisted changing her approach.

It was only when Mr. Patel's daughter, Priya, a nutrition student, joined the discussion, that a breakthrough occurred. They crafted a culturally competent plan, incorporating traditional Indian ingredients and cooking methods.

Mr. Patel felt more in control of the diabetes, and was energized by seeing his better blood sugar control, appreciating this new direction in his treatment plan.

Introduction:

Cultural competence is an essential component of evidence-based practice (EBP) in healthcare, as it ensures that healthcare providers can effectively address the unique needs of diverse patient populations. Within the realm of biomedical research related to EBP, it is important to note that existing studies often use samples that do not accurately represent real patient populations, often neglecting factors like the presence of comorbidities or demographic factors. Furthermore, the scope and duration of outcome measurements are typically limited, failing to account for long-term consequences. This lack of representativeness poses a significant challenge when it comes to assessing the effectiveness of treatments for specific populations.

When we extend this critique to encompass culturally competent medicine (CCM), the existing EBP literature falls short in adequately representing the cultural diversity of the population. Studies conducted with specific groups are frequently generalized to real-life situations (4). Much of the research primarily focuses on Western, middle-class, educated individuals while under-representing ethnocultural groups that constitute a substantial portion of potential patients.

Current Limitations on EBP and Cultural Competence:

While the evidence-based practice (EBP) movement strives to standardize clinical procedures based on empirical evidence, multiculturalist advocates express concerns about potential cultural biases in medical knowledge and practice. They contend that mainstream therapeutic approaches may not be suitable for culturally diverse populations, emphasizing the importance of cultural competence in delivering services tailored to diverse patient groups (1). Accommodating cultural diversity within narrowly prescriptive clinical practices presents a challenge that calls for multidisciplinary discussions among healthcare providers, biomedical researchers, and social scientists to expand the application of EBP beyond the cultural mainstream.

EBM demands the integration of the best evidence available with clinicians' expertise and patient's unique values and circumstances. CCM traditionally took a stronger role in prioritizing individual patient needs and cultural considerations. However, recent developments have witnessed EBM incorporating aspects of patient-centered care and individual preferences, while CCM has integrated methods to prevent cultural stereotyping and oversimplification. These trends open the door to synergies between the two approaches, potentially leading to a more unified framework for EBM and CCM to complement each other (2).

Identifying the Discord:

Shifting our focus to the realm of research, this section explores how cultural considerations influence study design and data interpretation, highlighting the essential role of cultural sensitivity in evidence-based practice.

Within healthcare, it is imperative to understand and address the nuances of culture. The traditional notion of cultural competence often oversimplifies culture, narrowing it down to ethno-racial identity. However, culture is profoundly intricate and ever evolving, encompassing intersubjective systems of meaning and practices intricately tied to specific social contexts. This oversimplification can perpetuate healthcare disparities, particularly but not only in the realm of mental health. Additionally, the interplay between EBP and CCM in healthcare adds a fascinating dimension. EBP relies on CCM to grasp the diversity of populations and make knowledge locally relevant, while conversely, CCM depends on EBP to validate its practices and adapt general knowledge into culturally appropriate interventions.

Despite the potential for mutual support, EBP and CCM are often independent and sometimes conflicting. EBP aspires to generate widely applicable knowledge but can fall short if studies fail to incorporate the nuances of culture. Striking the balance between respecting diversity while avoiding generalizations may present challenges when it comes to validating, replicating, or extending data to advance the field of medicine.

Synergies in Implementing Existing Data and Creating New Data:

To establish a solid foundation, it is crucial for all practitioners to recognize that despite earnest efforts to apply evidence and guidelines, healthcare practices are profoundly shaped by cultural, political, and economic factors. The translation of research evidence into clinical practice is a multifaceted process often entailing value-based decisions. Proponents of EBP acknowledge the importance of integrating scientific evidence with patients' values and life contexts. Nevertheless, the seamless integration of diverse knowledge sources remains a relatively unexplored and underdeveloped area.

The influence of EBP extends to the very conception of treatment goals and priorities. This approach sometimes pathologizes experiences, particularly relevant to psychiatric diagnoses but applicable to medical diagnoses as well, that might be considered normal in different cultural contexts (1). Changes in diagnostic categories and treatment approaches are influenced by cultural processes, marketing, resources and appeals to scientific evidence.

Integrating EBP and cultural competence in healthcare necessitates recognizing various types of knowledge, measuring a broader array of outcomes over extended time frames, and addressing cognitive biases and cultural values. Achieving this integration calls for methodological, epistemological, and political pluralism. To address the limitations of a purely evidence-based medicine (EBM)-based model, it is worth contemplating a shift towards person-centered applications within EBM. This approach emphasizes understanding illness and designing interventions based on patients' and families' perspectives, harnessing their strengths and resources. However, implementing this shift faces epistemological challenges and the reconciliation of varying knowledge claims.

Acknowledging the significance of multicultural ways of knowing and healing is imperative. Different communities possess unique epistemologies, ontologies, and sources of authority intertwined with their cultural identity and worldviews. Their healing objectives operate on various levels, including the sociopolitical, communal, and cognitive-emotional. To assess the effectiveness of these practices, one must consider different mechanisms and outcomes.

Bridging the gap between distinct knowledge frameworks and epistemic communities requires shared epistemic assumptions and a respectful dialogue that recognizes the diverse goals and outcomes associated with different epistemic perspectives. Understanding that epistemic and cultural differences can significantly influence healthcare interventions and outcomes underscores the necessity for methodological pluralism and, ultimately, the emergence of new forms of political recognition and engagement.

Bringing the Theory to the Daily Practice in the Clinic and Within Research:

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

In the sections above, we've delved into academic-level philosophical discussions regarding the broader systems of knowledge encompassing evidence-based practice and culturally competent medicine. However, the heart of the matter lies in finding synergy within everyday clinical practice. As medical students, residents, fellows, and attending physicians, our quest to apply various clinical studies to patient care necessitates a 'mental checklist' of critical questions. One pressing concern is whether the emphasis on self-identified racial categories in clinical studies inadvertently reinforces the notion of 'biological diversity' within patient populations, thus solidifying race as a biological difference, despite its status as a social construct. Furthermore, we must reflect on whether our focus on including self-reported diverse patients in studies might be perceived as mere 'virtue signaling' or a check-box exercise to make us feel good, rather than addressing the deeper social determinants that fall under racial categories. Perhaps, it is equally important to consider factors like income brackets, zip codes, and others as stand-alone demographic categories, serving as more precise proxies for the intricate social determinants of health that are intertwined with race.

The most effective synergy between evidence-based practice (EBP) and culturally competent medicine (CCM) begins with a comprehensive understanding of the patient, including their context within the predominant culture, often White American culture, and their specific care goals. By prioritizing this patient-centered approach and actively engaging in open dialogue, healthcare providers can build a foundation of trust and mutual understanding. In this context, EBP can be thoughtfully applied through a shared-decision model, ensuring that treatment plans are not only rooted in the best available evidence but are also tailored to align with the patient's cultural beliefs, values, and preferences.

The intricate interplay between culture and healthcare practices underscores the necessity for a comprehensive, culturally sensitive approach that embraces the synergy between Evidence-Based Practice and Cultural Competence. A nuanced understanding of culture and a harmonious blending of diverse knowledge frameworks hold the promise of improved healthcare outcomes and greater equity in healthcare delivery.

Dr. Amir Khan, a caring physician, had a patient named Fatima who needed to take medication with food, but she was fasting during Ramadan. Dr. Khan recognized the significance of her religious observance and worked with her and the community Imam to find a solution.

They adjusted the timing of Fatima's medication to coincide with her pre-dawn meal (Suhoor), allowing her to fast while adhering to her health regimen. Dr. Khan's cultural competence and flexibility ensured Fatima's religious and medical needs were met during Ramadan.

References:

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

(1) Gone JP. Reconciling evidence-based practice and cultural competence in mental health services: introduction to a special issue. *Transcult Psychiatry*. 2015 Apr;52(2):139-49. doi: 10.1177/1363461514568239. PMID: 25808532.

(2) Hasnain-Wynia R, Pierce D, Wynia M, Johnson M. Practicing Evidence-Based and Culturally Competent Medicine: Is it Possible? *Virtual Mentor*. 2007 Aug 1;9(8):572-8. doi: 10.1001/virtualmentor.2007.9.8.oped1-0708. PMID: 23218153.

(3) Huey SJ Jr, Tilley JL, Jones EO, Smith CA. The contribution of cultural competence to evidence-based care for ethnically diverse populations. *Annu Rev Clin Psychol*. 2014;10:305-38. doi: 10.1146/annurev-clinpsy-032813-153729. Epub 2014 Jan 15. PMID: 24437436.

(4) Kirmayer LJ. Cultural competence and evidence-based practice in mental health: epistemic communities and the politics of pluralism. *Soc Sci Med*. 2012 Jul;75(2):249-56. doi: 10.1016/j.socscimed.2012.03.018. Epub 2012 Apr 23. PMID: 22575699.

Submitted 11-4-2023

I.4 Clinical Decision Making, Gut Feeling or Hard Rules? (Jacob Markwood, GSM4)

The delivery of safe and effective care requires the clinician to have a solid process of making the right clinical decision in the pressure of complex patient encounters. There has been much research on how people make decisions when it matters most, traps that everyone can fall prey to, and tools to improve this everyday process.

A notable work on this topic is *Thinking Fast and Slow* by Daniel Kahneman where he describes the two ways that the human brain can process and interpret information to make the best decision. He proposes that the human brain can use System 1 which is the fast, seemingly automatic response to input, compared to System 2 which is the methodical and calculated evaluation of the data. It is usually held that System 2 thinking is less error prone given the increased attention to detail compared to System 1 which is often compared to “gut instinct”.¹ The key to medical decision making is striking the right balance of accuracy and efficiency in the time-constrained but high stakes environment which is modern medical practice. An increasingly popular approach to addressing this conundrum is the creation and utilization of clinical decision rules (CDRs).

The goal of CDRs is to improve consistency and confidence in making a diagnosis or deciding on a course for further testing or treatment given the patient’s symptoms and baseline characteristics. Like all evidence-based medicine, CDRs can only be trusted if the methods and interpretation of the research data are consistent and accurate. CDRs are developed out of a research question. These questions usually involve the following format, “Given a certain patient presentation with certain baseline characteristics such as age, sex, risk factors, plus or minus certain test results, what is the likelihood that this patient has the diagnosis in question”?²

To put this in the language of CDR development, first there needs to be *defined outcome*, for example does this patient have a pulmonary embolism? The next step is what are the *predictor variables* that can be used if present or absent to predict if the outcome of interest is more or less likely. The development or *derivation phase* of a CDR involves the collection and analysis of the most accurate and consistent predictors to determine the outcome in question. Once an initial model has been developed the validation phase requires the broad application of the rule to determine if the predictors hold up in real world clinical practice with a sufficient level of sensitivity and specificity to make it useful.² Finally the third stage is the implementation phase where the question of whether the

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

intended rule is having the hoped for and anticipated consequence on the frequency of invasive testing or patient outcomes.³

In the complexity of medical decision making having a “rule” to fall back on can be a comfort to the clinician when the best decision is not clear, or the stakes are high. While this sense of a having definitive answer is appealing, the outcomes of CDRs may not always be superior to clinical judgment or “gestalt”.⁴ One reason for this is that at their core CDRs require varying degrees of clinical judgment to determine if the rule applies to a particular patient or if a certain *predictor variable* is present or not. Morgenstern makes the astute observation, that if a symptom is binary with little room for variable interpretation a CDR holds up better than if the presentation is complex and there is room for interpretation of the *predictor variables*.⁴ For example, the Ottawa Ankle rule, used to determine if a patient presenting with an ankle injury or pain should get an X-ray has generally outperformed a clinician’s judgment because it is a straightforward question of the presence of absence of an ankle fracture. Also, the predictors are simple, but even still they can be subjective, including pain at certain locations of the foot and ankle, and a patient’s ability to bear weight. With a good clinical exam and training in application, this CDR has reduced unnecessary X-rays.⁴ However, when CDRs are applied to more complex clinical decisions with more room for individual interpretation and clinical variability, they are more error prone. For example, In the Well’s Criteria for Pulmonary Embolism (PE) one of the questions which can add or eliminate 3 out of 12.5 points toward the likelihood that a patient has a PE and what further testing is required: “Is PE the #1 diagnosis OR equally likely?”⁵ The answer to this question will inevitably vary based on the experience of clinician, recent cases, and inherent biases.

Given the variability and complexity of clinical care, trying to fit a patient complaint in a CDR box may not be as failsafe as we hoped, even though having the CDR app on our phone, and putting rule names like PERC, HEART, PECARN, and NEXUS in our charts make us feel more certain than is justified. In 2017 Schriger et al. completed a survey of the efficacy of clinical decision aids and rules compared to physician clinical judgment alone. First, they found that CDRs were infrequently compared to clinical judgment and in the 21 studies that did only 2 suggested that the CDR was better than clinical judgment alone. Given the results of their findings, they made the striking conclusion, “Just as we should not introduce a new medical treatment until there is evidence from well- designed studies that it outperforms current therapy so also, we should not advocate clinical decision aids...until they are proven superior to physician judgment”.⁶

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Even if not all CDRs are superior to “clinician gestalt”, Morgenstern presents a valid comment that the research involved in creating and validating a CDR is impressive.⁴ Stiell outlines the complexity and rigor of developing the Canadian C-Spine Rule which attempts to answer the question of whether a stable trauma patient requires imaging to ‘clear’ the cervical spine. The initial study to look at this question involved 8,924 patients, validated in a study of 8,283 patients, and 11,648 patients were included in an implementation style analysis; this is close to 30,000 patient cases to answer a seemingly simple but important clinical question.^{3, 4}

CDRs are frequently used and are valuable adjuncts when an answer is not clear, and the clinician’s bandwidth is limited. They are also an excellent example of the need for thorough investigation into the evidence when making clinical decisions.

References:

1. Denham. (2012). Thinking, fast and slow. *Journal of Communication.*, 62(5).
<https://doi.org/info:doi/>
2. Clinical Decision Rules. (2012). In *Evidence-Based Emergency Care* (pp. 36–43). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118482117.ch4>
3. Stiell, I. G., & Bennett, C. (2007). Implementation of Clinical Decision Rules in the Emergency Department. *Academic Emergency Medicine*, 14(11), 955–959.
<https://doi.org/10.1197/j.aem.2007.06.039>
4. Morgenstern, J. Clinical decision rules are ruining medicine, First10EM, February 2, 2023.
Available at: <https://doi.org/10.51684/FIRS.129162>
5. <https://www.mdcalc.com/calc/115/wells-criteria-pulmonary-embolism>
6. Schriger, D. L., Elder, J. W., & Cooper, R. J. (2017). Structured Clinical Decision Aids Are Seldom Compared With Subjective Physician Judgment, and are Seldom Superior. *Annals of Emergency Medicine*, 70(3), 338-344.e3. <https://doi.org/10.1016/j.annemergmed.2016.12.004>

Submitted 2/10/2024

Section II. Types of Studies

II.1 A Primer on the Design of Studies – (Jacqueline Raicek)

Study Designs in Medicine

1. Basic studies

- a. Animal studies
- b. Method development
- c. Genetic
- d. Cell

Investigate the cause-outcome relationships between a dependent variable and independent variable, such as animal experiment, genetic and cell studies. Method development studies investigate the development and improvement of biochemical, imaging, and biometric methods.

2. Observational studies

- a. Descriptive
 - i. Case report
 - ii. Case series
 - iii. Cross-sectional (descriptive or prevalence)
- b. Analytical
 - i. Cross-sectional, survey
 - ii. Case-control
 - iii. Cohort

Describes what is happening in a population, for example, the prevalence, incidence, or experience of a group. Often the first step or initial inquiry into a new topic, event, disease, or condition.

Attempts to quantify the relationship between two factors, effect of an intervention or exposure on an outcome.

3. Experimental/Interventional studies

- a. Randomized controlled
- b. Non-randomized controlled
- c. Self-controlled
- d. Crossover

Compare the effect of treatments or interventions with control in humans. Placebo or different treatments or interventions may be used as controls. Designed to reduce bias.

4. Economic evaluations

- a. Cost analysis
- b. Cost-minimization analysis
- c. Cost-utility analysis
- d. Cost-effectiveness analysis
- e. Cost-benefit analysis

Evaluate total cost of disease or health condition on society; compare alternative intervention's cost and outcomes; evaluate cost and benefit of alternative interventions.

Meta-analysis combines the statistical results of different studies in a particular clinical area and

5. Meta-analysis (including Network Meta-analysis) and Systematic review

systematic reviews evaluates and interprets the evidence of all studies conducted in a clinical area.

Sourced from [Balkan Med J.](#) 2014 Dec;31(4):273-7 and Center for Evidence Based Medicine, University of Oxford (<https://www.cebm.net/2014/04/study-designs/>)

January 2019

II.2 Factorial Design, Main Effect, and Interactions (Yi Zhang)

You may have come across a 2x2 factorial design in your experience of reading research articles. What exactly is the structure of this design?

A factorial experimental design consists of factors and levels. A *factor* is an independent variable. Each factor has a certain number of *levels*. Let's take an example.

Let's say we wish to look at basketball players and see if certain factors affect how many points they score. We can start by looking at two independent variables, for example age and amount of pregame Gatorade. Because each independent variable is a factor, there are two factors.

Factor 1: age

Factor 2: Gatorade

For each factor, there can be different levels. Let's say we want to look at those that are age 10 and those that are age 15. We are choosing two levels for age. For the amount of Gatorade drunk before the game, let's choose 1, 2, and 3 cups. So, there are three levels for amount of Gatorade.

This is a 2x3 factorial design. The first slot refers to the first independent variable, age. The number "2" refers to the number of levels for age. The second slot refers to the second independent variable, amount of Gatorade. The number "3" refers to the number of levels for amount of Gatorade.

(A)x(B)x(C) ... etc.

Each parenthesis refers to an independent variable. A is the number of levels for the first independent variable. B is the number of levels for the second independent variable. C is the number of levels for the third independent variable, and so on.

Back to our example, if we added a third independent variable, shoe brand, with 4 levels (Nike, Adidas, New Balance, Asics), how would we express this?

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

This would be a 2x3x4 factorial design. The order of the independent variables is arbitrary. We could say 2x4x3, 3x2x4, 3x4x2, 4x3x2, 4x2x3, and it would all be referring to the same experiment.

Going back to the 2x3 design, we can also find the number of conditions by multiplying the numbers together. Therefore, there are 6 conditions.

Experimental Condition #	Age (years)	Gatorade before game (cups)
1	10	1
2	10	2
3	10	3
4	15	1
5	15	2
6	15	3

With a 2x3x4 design, there would be 24 conditions. In the example above, age is predetermined. In a randomized controlled trial, patients would be randomly assigned to these conditions.

Now let's discuss main effect, which looks at the effects of an independent variable on a dependent variable. If we wanted to look at the main effect of age on points scored, we would look at the data as if the other independent variable, amount of Gatorade, did not exist. And vice versa to look at the main effect of amount of Gatorade. Using the same pool of patients, we can look at the effect of multiple independent variables. This is one of the benefits to using a factorial design.

	1 cup Gatorade	2 cups Gatorade	3 cups Gatorade	
10 Years Old	10 points	15 points	20 points	Mean = 15 points
15 Years Old	20 points	25 points	30 points	Mean = 25 points
	Mean = 15 pts	Mean = 20 pts	Mean = 25 pts	

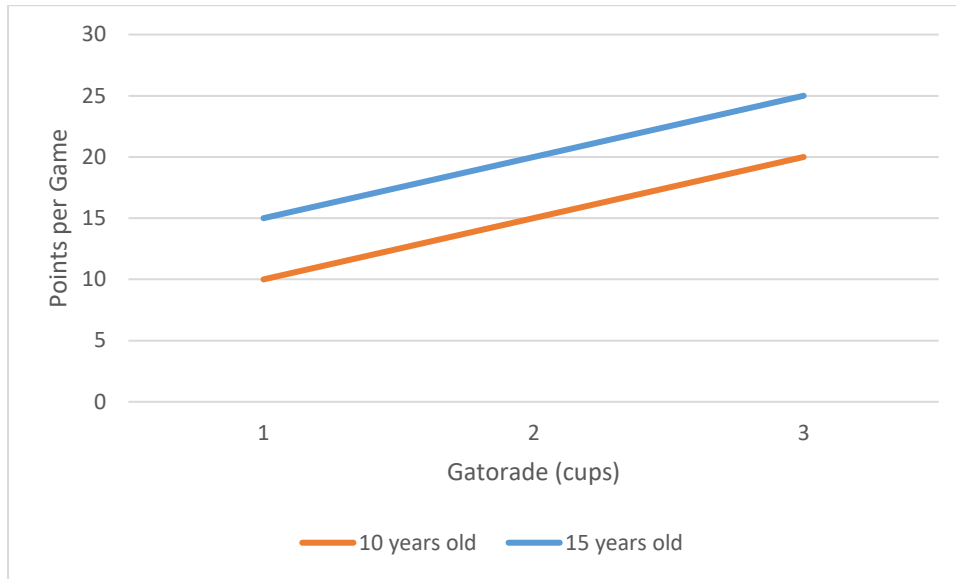
The above table lists the data for the dependent variable, points per game, in relation to the two independent variables in our hypothetical experiment.

We can see that the main effect of 1 cup of Gatorade is 5 points per game. This is the same regardless of whether we are looking at 10-year-olds or 15-year-olds. Essentially, this is looking at the effect of one independent variable on the dependent variable of interest. If we were to take away the age stratification and just look at the means, it would be 15 pts for 1 cup, 20 pts for 2 cups, and 25 pts for 3 cups Gatorade, with the main effect of 5 pts. We are used to this- many studies look at just one independent variable and one dependent variable.

Likewise, the main effect of age, in this case 5 years, is 10 points. This is the same regardless of the amount of pre-game Gatorade.

The effect of Gatorade is the same from 1 to 2 cups, from 2 to 3 cups, and regardless of age. The effect of age is the same regardless of cups of Gatorade. This means there are **no interactions**.

Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine



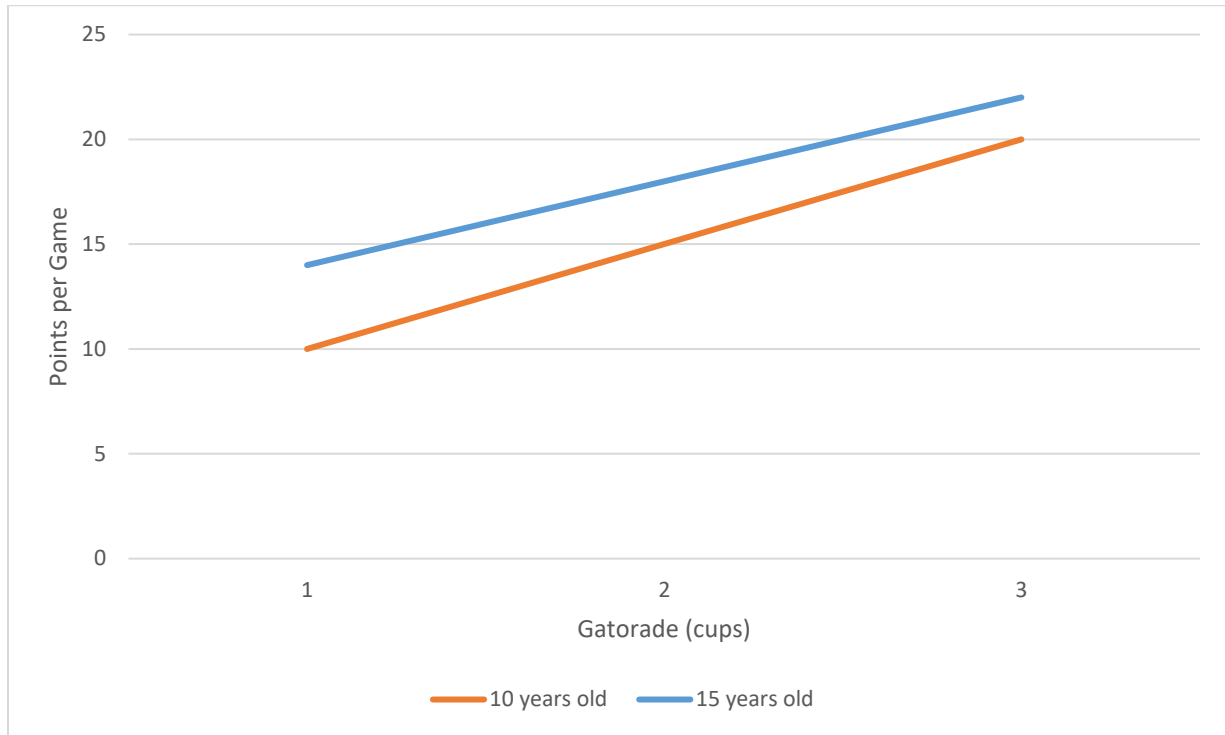
This is a graphical representation of the data. **When the lines are parallel, there are no interactions.**

Now, let's say the data were different.

	1 cup Gatorade	2 cups Gatorade	3 cups Gatorade	
10 Years Old	10 points	15 points	20 points	Mean = 15 points
15 Years Old	14 points	18 points	22 points	Mean = 18 points
	Mean = 12 pts	Mean = 16.5 pts	Mean = 21 pts	

Here, we see that the effect of 1 cup of Gatorade is 5 points in 10-year-olds, and 4 points in 15-year-olds. Also, the effect of age (age 15 compared to age 10) is 4 points with 1 cup Gatorade, 3 points with 2 cups Gatorade, and 2 points with 3 cups Gatorade. Since the effect is not uniform all the way across, that means there is an interaction.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine



On the graph, we see that the lines are not parallel, and would meet at some point if the data were to extend further in one direction or another. This means that there is an interaction. Something about age changes the effect that Gatorade has.

How does this apply to clinical trials? Let's take the example of a 2x2 factorial design. This means there are 4 conditions. Let's say we wanted to look at aspirin and apixaban (independent variables) and incidence of major bleeds (dependent variable). The numbers in the following table refer to the number of patients in each group.

	Aspirin	Placebo	Total
Apixaban	100	100	200
Placebo	100	100	200
Total	200	200	400

There is a total of 400 patients, with 200 randomized to aspirin and 200 to placebo, as well as 200 to apixaban and 200 to placebo. With these 400 patients, we are essentially conducting two parallel trials: aspirin vs. placebo, and apixaban vs. placebo. In addition, we can compare aspirin and apixaban together, to aspirin alone, apixaban alone, and placebo. We can look into how aspirin and apixaban work together and see if there are any interactions. This is the advantage to using a factorial design. One would need to make sure to sufficiently power the study for each of the four conditions in this hypothetical study.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Hopefully by now you have a better understanding of factorial design, main effect, and interactions, and how this can apply to clinical trials as well as examples outside of clinical trials. In summary, when choosing an experimental design, one important consideration is which one delivers the most statistical power with the fewest subjects. If the research questions call for direct comparison of individual experimental conditions, as is required when treatment packages are being compared, then this design will usually be an RCT. If the research questions call for assessing the effects of individual components of an intervention, then this design will usually be a factorial experiment.

References:

1. <https://www.methodology.psu.edu/ra/most/factorial/>
2. https://www.youtube.com/watch?v=rwQYLtG_AYI

Submitted 10-17-2020

II.3 Randomized Controlled Trial (RCT) Design: Key Elements in the Gold Standard of EBM (Keegan O’Hern)

Overview:

This chapter aims to introduce the design process of a randomized controlled trial (RCT). It will discuss the fundamentals and significance of forming a clinical question, randomization, blinding, bias, and statistical analysis (sample size and power calculations). This chapter is supplemented by the works of other chapters on these topics in the Evidence Based Medicine Elective Guide and acts to unify many of these topics in one narrative review.

Introduction:

RCTs are not only Dr. Ross’ favorite trial design but are one of the quintessential tools in evidence-based medicine in that they are designed to directly answer a clinical question. While other study types, such as case studies, case series, cohort studies, and the like raise important questions, they are insufficient to prove causality. The RCT forms two identical groups and attempts to control as many variables as possible, and introduces an intervention (e.g., therapy) to isolate its effect on the outcome of interest. To tackle this endeavor, one must first understand what types of questions can be answered by an RCT.

PICO Questions:

Anyone who has taken the EBM For Life! Elective should know what “PICO” stands for: Patient(s), Intervention(s), Comparison (Control), Outcome(s). It defines plainly for the research team, and the audience, who are the key players in answering a question, what you aim to do with said players, what your control group is for comparison, and by what measure(s) you deem to ascertain the effect of the proposed intervention. While this is of utmost importance in designing an RCT, it has been discussed in prior chapters and throughout this course, but each step is critical in designing an RCT. I would like to discuss some of the intricacies of the last three components, as they are the key to an RCT.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

The intervention is (usually) the entire point of an RCT – does an intervention lead to better or worse outcomes for the two groups? The most basic design has two arms: the intervention of choice, and a matched placebo for Comparison. However, it is not always ethical to give patients a placebo without adequate treatment (such as not treating severe atopic dermatitis in the placebo group while the intervention group derives a potential benefit from that arm). Thus, determining what the exact intervention is for each group, including the placebo and any adjunct therapies, is tantamount in determining if the benefit is from the intervention, the placebo, or a confounding variable. The placebo or comparison arm may not always have a completely inefficacious placebo, though each group should have similar demographics and disease severity (as typically seen in “Table 1”). Outcomes can make or break an RCT, because the primary (or secondary and beyond) outcome determines whether or not clinicians can distinguish a difference between the two groups. For diseases with objective (dichotomous) outcomes, the measurement is quite clear: e.g., did the patient live or die? Was there complete clearance of the tumor/lesion, or no? RCTs become tricky when there are no validated tools to assess a change, as when the outcome is subjective and requires some ingenuity to determine if there is an effect; this is common in dermatology where the determination of whether partial resolution of a lesion has occurred is up to the clinician looking at the photograph/patient. All of this to say that being able to come up with a concrete PICO statement is the first step of any RCT design.

Randomization:

A clinical trial is not an RCT if it lacks the R: randomization. Randomization is the gold standard of trial designs. Researchers aim to remain equipoise at the start of a trial: though they may have an inclination that a treatment works based on weaker literature, they generally are uncertain if the intervention will lead to net benefit or harm in the population of question (or else, why do the study at all?). To remove bias of the clinician selecting which patients go to the intervention or comparison group, randomization is key. The Consolidated Standards of Reporting Trials (CONSORT) guidelines that all RCTs are apt to follow says that groups in a trial should be formed by chance; it allows a statistical diversification of each group and “controls” for the chance that one group is inherently different than another. However, it does not always go according to plan, and a quick glance at the “Table 1” in most RCTs is warranted to ensure that RCTs are in fact randomized. Randomization at the individual level is standardized, but may also be done by groups, and there are some circumstances in which randomization may be unethical (a patient with severe, treatment-refractory disease being randomized to placebo is one example).

How are patients actually randomized? Simple randomization implies random number generation or a coin toss, but one may employ stratification to account for (e.g., control for) potential confounding variables like age, sex, etc., though this should be used sparingly for only characteristics/variables that may affect the outcome. Choosing the “right” randomization process requires that one understand how many arms will be utilized in the trial, as it would be difficult to randomize to three arms with a coin flip. It also depends on the calculated sample size needed to demonstrate a difference between the study arms, but more on that later. Randomization and blinding are inherent to a good study design, but the correct process for each is usually dependent on the context of the question at hand and what level of prior knowledge is present.

Bias in RCTs:

Bias is inherent in all study designs, and researchers must employ multiple means to reduce this entity to ensure the results are not impacted by confounders. Bias implies a systematic error, and thus is not random; the act of randomization as above helps limit these confounders due to randomness. However, processes like selection bias wherein participants are allocated to the intervention or placebo group based on baseline characteristics such as disease severity or age may impact the outcome of trials; this again speaks to the importance of randomization. Blinding similarly ensures that the researchers and patients do not change their expectations, or subconsciously their actions, for each arm of the trial. Finally, attrition bias may impact results based on the withdrawal of participants from a group, and this highlights the importance of an intention-to-treat analysis whereby participants are evaluated in the groups they were originally assigned to. Of note, expected withdrawal rate should be considered when determining your sample size (see below), as well as the ethical basis of the investigation at hand as there should not be an enormous attrition rate observed due to safety concerns, adverse effects, or lack of efficacy. There are a large number of biases that RCTs are designed to control for, and while the list of potential biases is too large to discuss here, the RCT is considered the gold standard for determining the difference between interventions and control given the efforts to minimize bias intrinsic to the RCT design model.

Phases of RCTs:

RCTs come in many shapes and flavors, though a tiered system helps readers understand what level of evidence is present and which questions are being asked in each phase of the trial. Phase 1 trials are just the very beginning of the RCT process with generally a smaller sample size to determine feasibility. Feasibility or proof of concept trials help determine whether larger trials will be useful: in a small group, does there appear to be some relative safety and efficacy? They also help you get the kinks out for later trials, as you learn whether you can (or cannot) run your trial as initially planned. Phase I trials are where the rubber meets the road and can make or break an RCT moving forward. Phase II commonly determine efficacy, but are not large enough to determine safety and effectiveness, that's where phase III comes in to determine the external validity of the intervention: does it work in the real world? Phase IV is usually after the intervention has been approved (after market studies) and will not be covered here as this is about starting an RCT from scratch.

Statistical Analysis and the “Myth” of the Significant p-Value:

Though an RCT can be performed to determine superiority of one intervention over another, the null hypothesis present for any two comparisons is that there is no difference between the intervention and placebo. In a head-to-head trial, the goal is to demonstrate that a health benefit (or harm) is obtained from the intervention, that is, that the estimate of efficacy lies above the control. While there are many ways to mathematically determine this, a point estimate—or mean value in the case of a continuous variable—aims to demonstrate that the “true value” of the measurement lies above or below the 95% confidence interval (CI), such that there is a 95% chance that there is a true difference between the measured outcome in the intervention and the control group. A non-inferiority or equivalence trial aims to show no difference between the intervention and control (often the control here is an approved treatment or intervention for the condition of interest), and statistically this is demonstrated by overlap in the 95% CI.

As has been covered elsewhere in this guide, the p value represents the chance that the results observed occurred due to chance, and it is commonly accepted that a value ≤ 0.05 is statistically significant, and those studies able to manipulate complex regressions to obtain a p value ≤ 0.05 often results in publication. It is important to note that values ≤ 0.1 still have a 90% chance of being observed outside of chance which may be significant depending on what is being tested. Additionally, the more that something is tested (e.g., large sample size in trials), the more likely one will find a statistically significant result, but an astute reader should determine the clinical significance of the results. One cannot assume that a p value ≤ 0.05 is meaningful, the context in which it was derived is critical to making meaning out of the numbers. A more nuanced and useful metric is the confidence interval, which gives one a range of possible true values and identifies with 95% confidence which range is likely to be true. CIs are attached to important metrics such as the relative risk reduction or increase, the absolute risk reduction or increase and the NNT (number to be treated to prevent one outcome) or NNH (number needed to cause a harmful outcome).

Sample Size and Power Calculations:

When designing an RCT, one must calculate what sample sizes would be required to obtain a statistically significant result based on the measurements and outcomes being obtained. One tool that has been developed by the Centers for Disease Control and Prevention (CDC) is OpenEpi (https://www.openepi.com/Menu/OE_Menu.htm) that allows one to use various calculators to determine appropriate sample sizes, make power calculations, and perform statistical analyses like ANOVA and t tests. These tools can be helpful in determining how large a sample size should be in addition to reading similar trials and determining how sample sizes were calculated.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

To determine a sample size, one needs to consider the effect size, p value, and power. While these all have mathematical derivations, they are concepts that can generally be asked as simple questions that one designing an RCT should consider: How large and in what direction (positive or negative) do the designers believe the difference between the two (or more) arms should be, and is that clinically meaningful? Of note, a small difference in the measured outcomes (termed effect size) will require a larger sample size, but more on that below. If the variables being measured have a large variation, the “error bars” can be quite large unless the sample size is large enough to help narrow the 95% CI, such that if variability is small, then the resulting sample size required will be small.

Often, the effect size of the intervention can be taken into account to determine how large a sample size should be to see a certain magnitude of difference between groups. Dartmouth’s Synergy biostatistician group is a helpful resource in making these calculations, and there is commonly a team of biostatisticians that play a pivotal role in helping to design and implement these aspects of RCTs, to adjust for participant attrition, and to interpret the results. As above, a p value of 0.05 is commonly used. Power (denoted as the Greek letter “beta”) is often set to 0.8-0.9 and demonstrates that if there is a difference between the groups, the trial has a large enough sample size to detect that difference. Being “under-powered” may be a problem for smaller trials such that an effect that is present is not observed due to the small sample size.

The complexity of statistical analysis is out of scope for this topic, but I point readers to the above resources as well as the references of this chapter for additional information.

Conclusion:

RCTs are considered the gold standard in evidence-based medicine and designing one from scratch requires an intimate knowledge of the components of its methodology and design. All good studies begin with a well-outlined question, and by understanding how an RCT is designed, we can all begin to improve our ability to interpret and create evidence-based medicine.

References:

1. Bhide A, Shah PS, Acharya G. A simplified guide to randomized controlled trials. *Acta Obstet Gynecol Scand.* 2018;97(4):380-387. doi:10.1111/aogs.13309
2. Parmar MKB, Sydes MR, Morris TP. How do you design randomized trials for smaller populations? A framework. *BMC Med.* 2016;14(1):183. doi:10.1186/s12916-016-0722-3
3. Brocklehurst P, Hoare Z. How to design a randomized controlled trial. *Br Dent J.* 2017;222(9):721-726. doi:10.1038/sj.bdj.2017.411

Submitted 10-9-2020

II.4 RCTs: Strengths and Limitations (Mariah Evarts)

Strengths of RCTs:

In *Evidence-Based Medicine: How to Practice and Teach EBM*, Straus et al state that “evidence-based medicine requires the integration of the best research evidence with our clinical expertise and our patient’s unique values and circumstances.” In order to fulfill the “best research evidence” component of practicing EBM, there must be an evaluation of strength of evidence, largely based on study design and implementation. The evidence hierarchy for testing treatment strongly favors the randomized controlled trial (RCT) as a design that uses randomization to reduce various sources of bias that plague observational and non-randomized studies. When used to test therapies, RCTs can draw causal links between an intervention and an outcome more easily than other studies because of the randomization and resulting lack of confounding bias.

What are the weaknesses of RCTs?

As much as RCTs are heralded as the gold standard for determining efficacy of one treatment as compared to another, they are not without limitations. Discussing weaknesses of RCTs is particularly important because it allows for a more nuanced understanding of reported findings.

Randomization:

A central component of RCTs is that recruited and eligible patients are randomized into treatment groups. Computer-generated randomization is often favored. On a conceptual level, randomization combats confounding bias, particularly confounders that have not been identified by the treatment team. In practice, however, there are a variety of ways that this can generate biased results. Schulz et al found that many studies had treatment groups that were statistically much more similar than chance would have predicted. The authors deem this to be due to “nonrandom allocation” rather than “replacement randomization”.

Exclusionary Criteria:

To decrease heterogeneity in treatment groups, RCTs often define a narrow set of selection criteria that may exclude important populations, including women and those over 65 years old. Likewise, subjects with comorbidities are often excluded. From a study design perspective, this increases internal validity which can be used as a marker for low risk of bias. From a clinical perspective, however, this obviously will decrease the applicability of the information, thus decreasing its external validity. Additionally, it is possible that by selecting for such a narrow group of subject characteristics, the study results may be artificially inflated.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Rothwell also reports on other ways that RCTs can select for certain types of patients even prior to randomization. One example is when study design uses a pre-randomization run-in period. All recruited subjects take a placebo and those subjects that are not adherent or have adverse events from the medication are excluded from the cohort to be randomized. Again, from a statistical perspective, reducing non-compliance would increase internal validity so that the therapy is truly being tested rather than testing a behavior (non-adherence) as well as the therapy. However, this may bias results but also leads to a non-representative sample, thus decreasing external validity.

Rothwell points out that these types of selection mechanisms are particularly troubling because they are often not reported and there is no quantifiable way to assess the external validity. He suggests that all trials report the number of eligible subjects that were not included in the randomization as well as the number of patients invited but who ultimately declined to participate in randomization.

Clinimetrics:

RCTs inherently need to have a measurable outcome within a reasonable time frame, thus studies often call for measuring either binary outcomes or indirect values such that a difficult-to-measure outcome can be quantified. There is a concern that this has led to a hyper-focus on the measurable and non-binary outcomes – effect on quality of life, distress, overall well-being, etc. – are given overall less value. Fava calls this a move to “clinimetrics” that de-emphasizes a biopsychosocial approach to medicine. True integration of evidence-based medicine can only occur when clinimetrics and the biopsychosocial approach can exist in the same space.

Variation Masked by Averaging:

As previously discussed, RCTs often have stringent inclusion and exclusion criteria as a method of increasing internal validity and potentially resulting in statistical significance which may not be replicated in the average person. Even if these criteria are loosened somewhat, however, there are resulting issues involving averaging across heterogeneity. For example, the averaging may conceal helpful information about a portion of participants that didn't respond to the therapy. This is a particular worry as the size of the study increases in participant number.

One suggestion that is commonly given in an attempt to deal with the heterogeneity is to create subgroups based on pathophysiological understanding and thus what subgroup of people may react differently to an intervention. This should never be done *post hoc* and direct conclusions from subgroup data are dangerous because studies are often not powered for that analysis. Regardless, it is tempting to analyze subgroup results and draw conclusions about why there are differences, potentially even extrapolating the information to a patient that matches the baseline characteristics of the subgroup more directly. *NEJM* and *JAMA* caution against this type of analysis and instead encourage the use of subgroup data to formulate thoughtful hypotheses for future studies.

References:

1. Fava, GA. “Evidence-based medicine was bound to fail: a report to Alvan Feinstein” *J Clin Epidemiol* 2017;84:3-7.

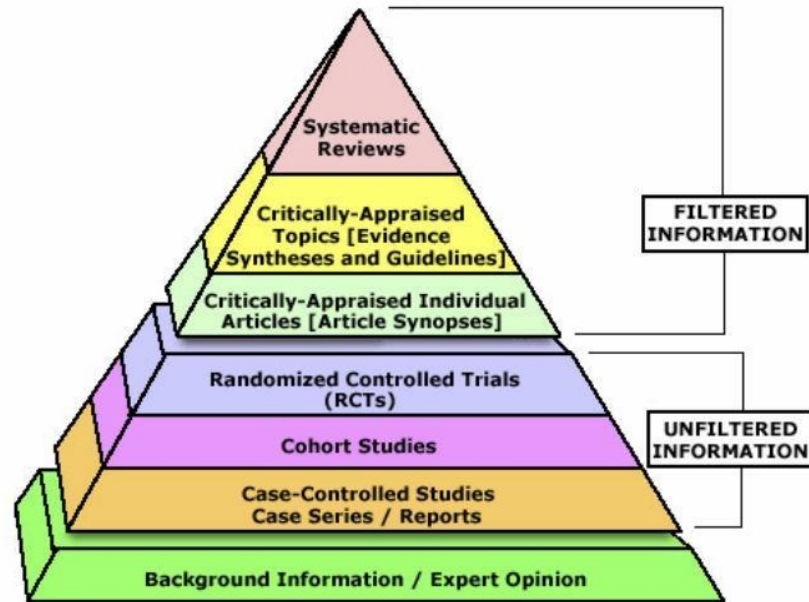
2. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. "Can treatment that is helpful on average be harmful to some patients?" *J Clin Epidemiol* 1996;49:395-400.
3. Nichol AD, et al. "Challenging issues in randomised controlled trials" *Injury* 2010;41S:S20-3.
4. Rothwell, PM. "External validity of randomised controlled trials: 'To whom do the results of this trial apply?'" *Lancet* 2005;365:82-93.
5. Schulz KF, et al. "Assessing the quality of randomization from reports of controlled trials published in obstetric and gynecology journals" *JAMA* 1994;272(2):125-8.
6. Wang, R et al. "Statistics in medicine--reporting of subgroup analyses in clinical trials" *N Engl J Med* 2007;357(21):2189-94.

Submitted- Feb 2019

II.5 Systematic Reviews and Meta-Analyses (Alex Donovan)

In the field of medicine, research advances are constantly guiding us in new directions regarding the way we evaluate, diagnose, and treat patients. Though we do our best to critically appraise each randomized control trial and decide how it will change our practice standards, the amount of information can be overwhelming, and even good RCTs contradict each other frequently. Fortunately, we have two forms of evidence-based literature that aim to synthesize the available evidence we do have, while taking the strengths and weaknesses of individual studies and synthesizing the results of the studies together to help generate more global conclusions, which can hopefully better guide our medical decision making.

Systematic Reviews and Meta-Analyses are described below.



Systematic review

A Systematic review is a summary of all available evidence meeting specific eligibility criteria that can be used to address a specific question. It is a synthesis and critical appraisal of all studies known to address the same specific research question with an aim to limit bias as much as possible.

A Systematic Review is the most transparent of reviews as it communicates methods and bias explicitly. The studies evaluated in a systematic review are selected very methodically based on specific criteria to answer the same scientific question and to minimize bias. These special reviews evaluate the differences in studies through a quantitative and qualitative means known as heterogeneity, which helps explain how similar or different the individual studies being compared are in order to help gauge how meaningful/applicable the results are. (Please see section 8 on Heterogeneity for more details.)

Meta-Analysis

A Meta-Analysis is a statistical method that combines the results from different studies to effectively provide more power than the individual studies alone. They are also helpful in synthesizing big-picture statistics such as incidence, prevalence, and diagnostic accuracy due to the larger numbers and ability to evaluate for trend among the individual studies, which cannot be seen in the individual studies themselves. Systematic reviews often include a meta-analysis to help demonstrate statistical results of the studies included.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Most often, the odds ratio or relative risk are the metrics used to demonstrate relative effect in a meta-analysis. The pooled effects of the individual randomized control trials are often portrayed in the form of a Forest Plot which demonstrates the odds ratio of each study with a 95% confidence interval, in comparison to “no treatment effect” (the equivalent of an odds ratio of 1.0). Overall, a meta-analysis helps us evaluate the amount and strength of evidence available to answer a specific medical question.

References:

1. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med.* 1997; 126:376.
2. LeLorier J, Grégoire G, Benhaddad A, et al. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N Engl J Med.* 1997; 337:536.
3. Lindsay, Uman. Systmatic Reviews and Meta-Analyses. *J Can Acad Child Adolesc Psychiatry.* 2011. 20 (1): 57-59.

II.6 Cluster Randomized Controlled Trials: What are they, when can they be used, and what biases might they introduce? (Emmalynn Moore, GSM4)

In creating clinical trials, one of the major decisions study designers must make is a method for randomization. While many people are likely most familiar with individual randomization, in which each participant (or unit, in the case of pregnant people and their fetus) enrolled gets randomized one at a time to a given study arm, there are many methods for randomization, all of which have benefits and considerations. One of those is the concept of cluster randomization, in which the unit of randomization are groups or batches of participants. For example, in a study about an intervention in a primary care setting, all the patients who get care at a particular office may be randomized to one arm, and patients at another practice may be randomized to another.

I was first introduced to the concept of cluster randomization when reading a study in which two different methods of induction of labor were being compared on a single labor floor. To facilitate this, the study authors decided to use a version of cluster randomization in which a randomizer decided which protocol/method would be used at the start of each week, and all patients admitted in that week received that protocol¹. This was reset at the start of each week. The study authors picked this method in part because it was seen as more practical, as randomizing on the individual level seemed to have a high risk of patients inadvertently receiving the wrong protocol if staff was required to keep track of different methods for different patients simultaneously. As a reader, I thought this was an interesting way of formulating a study to try to imitate actual practice, and I wanted to learn more about cluster randomization and what we should think about as readers of evidence-based medicine when cluster randomization is used.

Cluster randomization was first introduced in the 1940's in trials evaluating education², in which all students in a classroom or school were randomized to the same arm, in part because it was felt to be more practical to implement education interventions at this level than to try to give each individual student different curriculums. Now, while cluster randomization continues to be popular in education studies, it has gained traction in the health care and public health sectors. Multiple studies show that cluster randomization has become especially more common in the last few decades², perhaps along with the rise in popularity of pragmatic trial designs, which attempt to closely mimic "real world" conditions to make study findings more applicable and generalizable to practice. In the healthcare research field, cluster randomization is often used to evaluate new patient care or protocols³, such as

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

the induction of labor study discussed above. Other examples include testing methods to promote increased hand washing by comparing methods on different units of a hospital or evaluating the effectiveness of two types of counseling on weight loss in a primary care setting by assigning all the patients of one physician to one intervention and the patients of another physician to the other. As in the induction of labor study, clusters can also be based on time² – for example, any patient being enrolled in the study in a given month being sent to the same intervention arm.

In addition to practicality, cluster randomization was introduced in an attempt to minimize contamination between participants of different intervention assignments. In its roots in educational studies, contamination seemed sure to impact outcomes if individual children in a classroom were randomized to different interventions, as it wasn't practical to separate the children, and as such, children in one arm would inevitably be exposed to some aspects of the intervention of the other arm, and vice versa. This is also an argument used in favor of cluster randomization in healthcare and public health studies. Consider the example of weight loss counseling in primary care: in theory, if the same physician gave different versions of counseling to her patients, those patients might encounter each other in a waiting room and discuss what they had heard from the doctor, potentially contaminating the interventions and leading to dilution bias (a version of Type II error in which we are more likely to accept the null hypothesis due to crossover between arms)². The idea of contamination is especially salient in the discussion of cluster randomization in clinical trials of vaccines. In theory, if participants are randomized at the individual level to vaccine or placebo, there could be an underestimate of the effectiveness of the vaccine, as those who received placebo would still be protected to some degree by herd immunity from being around those who received a vaccine². If participants are randomized in clusters, such as by primary care office or geographical area, there is a significantly lower chance that participants in placebo arms would be interacting with, and protected by the immunity of, those in the vaccine arms. These are just two examples of why avoiding contamination is so often cited as the primary reason for randomizing on a cluster level instead of an individual level in healthcare intervention studies.

While there are clearly benefits to randomization by cluster, it is important to remember that certain biases can be introduced with this method. One key consideration when evaluating cluster randomized trials is that they inherently require larger sample sizes to produce adequately powered results^{2,4}. In individually randomized trials, we assume there is no correlation between outcomes for individuals. However, randomization by cluster makes this assumption untrue – by the nature of the

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

randomization method, the outcomes that individuals have is correlated to their cluster, and as such, the variation we see between clusters may be less likely attributed to the interventions and more to features that are similar within clusters but vary between them². This concept of the amount of variation that can be attributed to features of the cluster is called the intra-cluster correlation coefficient (ICC)², and the higher the ICC, the more participants need to be recruited into the study to achieve the same power as would be required if the study used individual randomization. While this might make it practically more difficult to conduct a study with statistically powerful results, it also introduces an ethical question to consider: if the same study could be done with fewer participants using individual randomization, is it acceptable to expose more people to an intervention in order to use cluster randomization?⁴ In the discourse about cluster randomization, scholars suggest this question be strongly considered when deciding whether cluster randomization is justified.

Cluster randomized trials are also prone to selection bias. In many cluster randomized trials, the clusters are determined, and then participants from that potential cluster are recruited. Because the potential cluster must be known in order to accurately recruit from it, there is no real possibility of blinding to study arm assignment. Additionally, recruiters will already know what arm a potential participant from a given cluster will be assigned to if they agree to participate, which could lead to bias in who is ultimately recruited into the study^{2,5}. Selection bias can also come into play at the level of the cluster – if not done independently and randomly, clusters might be assigned to interventions for reasons that could potentially impact outcomes.

There are also some considerations when it comes to statistical analysis of cluster randomized studies. Authors must make sure to evaluate data at the level of the cluster instead of at the level of the individual, again because the assumption that outcomes between individuals are not correlate is untrue for cluster trials. Randomization at the cluster level followed by analysis at the individual level can lead to false conclusions, so researchers should use statistical approaches that allow evaluation at the cluster level. One example would be to use a two sample t-test using cluster means instead of individual values^{2,4}. Other more complicated methods also exist. Overall, when conducting cluster randomized trials, the correlation between outcomes within clusters must be considered at every step to try to ensure accurate and reliable results and conclusions.

It seems clear that there are various settings in which choosing cluster randomization over individual randomization provides benefits, both in practicality and in minimizing the potential impacts of contamination between study groups. However, it is important to remember that the use of cluster

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

randomization also introduces some potential biases as well as complications with statistical analysis. Knowing the pros and cons of cluster randomization can help us be more critical readers of evidence-based medicine when we encounter this method in studies we are evaluating. Moving forward, I know I feel more confident in discussing these types of trials, understanding the strengths and weakness of this type of trial design – hopefully this chapter helps you feel that way, too!

References:

1. Adhikari EH, McGuire J, Lo J, McIntire DD, Spong CY, Nelson DB. Vaginal Compared With Oral Misoprostol Induction at Term: A Cluster Randomized Controlled Trial. *Obstet Gynecol.* 2024 Feb 1;143(2):256-264.
2. Puffer S, Torgerson DJ, Watson J. Cluster randomized controlled trials. *J. Eval. Clin. Pract.* 2004 Sep 16; 11(5):479-483.
3. Heagerty PJ. Cluster Randomized Trials. *Rethinking Clinical Trials.* November 3, 2023. Accessed February 14, 2024. <https://rethinkingclinicaltrials.org/chapters/design/experimental-designs-and-randomization-schemes/cluster-randomized-trials/>.
4. Hemming K, Taljaard M. Key considerations for designing, conducting and analysing a cluster randomized trial. *International Journal of Epidemiology.* 2023;52(5):1648-1658.
doi:10.1093/ije/dyad064
5. Giraudeau B, Weijer C, Eldridge SM, Hemming K, Taljaard M. Why and when should we cluster randomize? *Journal of Epidemiology and Population Health.* 2024;72(1).
doi:10.1016/j.jep.2024.202197

Submitted 2-26-24

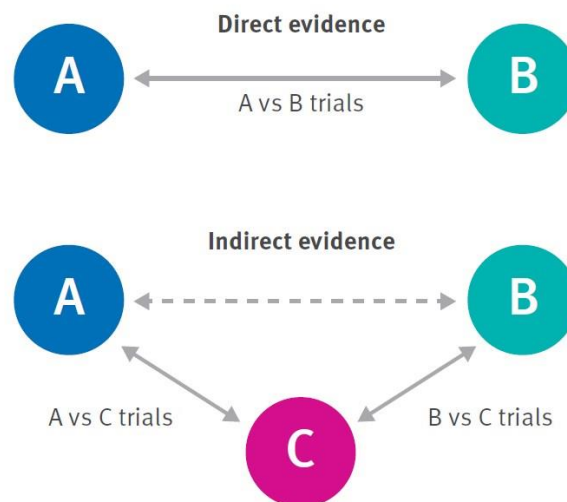
II.7 Network Meta-Analysis- Explanation and Interpretation of a Unique Tool for EBM (David Styren)

What are network meta-analyses?

This guide will not seek to explain the finer mechanics of *generating* a network meta-analysis (NMA) but will instead focus on *understanding and interpreting* NMAs. So, what are NMAs? Network meta-analysis is a specific subset of “standard” or “traditional” meta-analysis that is being utilized more and more frequently in systematic reviews. Simply put, NMAs are meta-analyses that, through advanced statistics and data interpretation, allow the indirect comparison of different interventions where head-to-head trials do not exist (or are limited).

You might ask yourself, “What does that mean in terms of real-world practice?” For example, say that a clinician is trying to decide between two drug therapies (Drug A or Drug B) for the treatment of one of his/her patients. Through a literature review, the clinician notices that numerous trials compare Drug A or Drug B to placebo, but no head-to-head randomized controlled trials exist (or perhaps only one or two small trials). There are even systematic reviews of both Drug A and Drug B to help separately determine the treatment effect of each therapy in a large patient population. Traditionally, the clinician in this situation would have to weigh the relative treatment effects of each drug in isolation and attempt to infer which drug may be the better choice for his/her patient. NMAs were developed to support clinicians in exactly this kind of scenario. NMAs utilize and synthesize data drawn from separate trials in order to mimic a head-to-head trial as closely as possible and provide a clinician increased confidence in making therapy decisions.

The theory behind network analysis is fairly straightforward, although the formation of a network analysis is anything but simple. Essentially, the supposition is that if Drug A and Drug B are both compared to a common comparator (usually placebo), and if their study designs and populations were similar, then via a version of the transitive property, A can be indirectly compared to B.



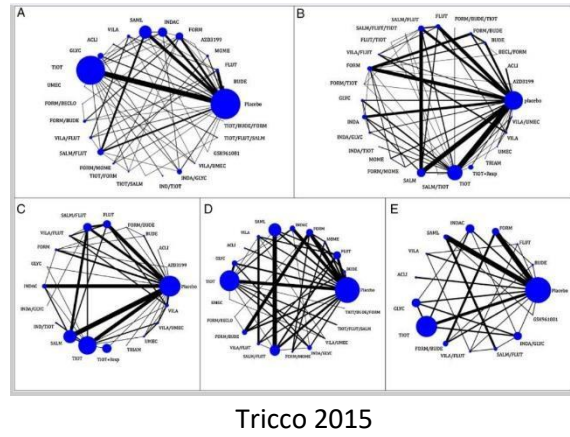
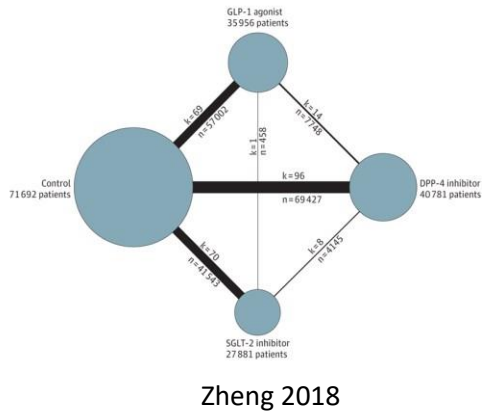
Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Riley 2017

The strength of the indirect comparisons relies heavily on the amount of data available for evaluation, as well as the quality of the studies involved, just as traditional meta-analyses do. A NMA that draws conclusions from very few studies, with few subjects, and questionable quality will be doubted just as a traditional meta-analysis with those limitations would. However, whereas a high-quality meta-analysis can usually only provide conclusions about a single comparison (Drug A versus placebo, etc.), network meta-analysis can compare Drug A to Drug B, AND Drug A to Drug C and Drug B to C and so on. Often, NMAs will include a visual depiction of the studies involved in their analysis (called network plots), illustrating how the studies relate to one another and the strength of the connection between the studies.

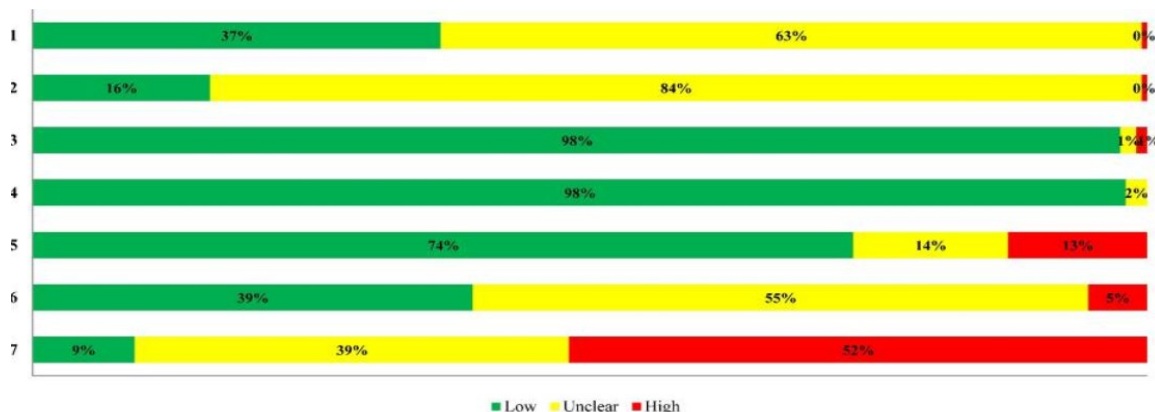
As illustrated in the examples below, network plots can be relatively simple or highly complex depending on the number of interventions being studied and the number of subjects/studies available. There are similar features to each network plot, however. In brief, each therapy being studied is usually labeled on the map, and lines are drawn between therapies to illustrate studies that have directly compared to the two therapies. Often each therapy will be connected to placebo (as that is the most common “common comparator”), but therapies can have lines drawn between one another when direct head-to-head studies have been performed. The number of patients that have been exposed to a particular therapy is represented as a circle of varying size, which increases proportionally to the number of patients. This is a useful feature as it allows the clinician to have perspective when interpreting conclusions later in the NMA. For example, if the NMA determines there is a significant benefit or harm associated with Drug A versus Drug B, but an extremely small patient population was exposed to Drug A, the relative imbalance in the two patient populations may have impacted the results. Another useful visual aid included in network maps is the thickness of the lines connecting certain therapies. In most network maps, with an increasing number of studies comparing two therapies, the thickness of the line increases. Similar to the size of the circles, the thicker lines indicate more studies have studied that particular comparison and can provide potentially stronger conclusions than a comparison where extremely few studies have been performed. Although the purpose of NMA is to provide additional data when a relative paucity of studies exist, just as with traditional meta-analysis, the greater the amount of data, the stronger the conclusions.

Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine



How are network meta-analyses made, and what makes them valid?

NMAs are designed much like any other, more “traditional”, meta-analyses. A clinical question is presented, and a literature review is performed to locate and assess any and all studies that might be useful in answering the initial clinical question. Often with NMAs, the clinical question involves multiple treatment modalities or options, and trials involving those options are broken down into two broad categories: head-to-head studies between treatments and treatment versus placebo studies. The vast majority of the time, there are more treatment versus placebo studies, and these are primarily what are used to develop the network meta-analysis. Information from each of the studies is carefully extracted and pooled with other data before being subjected to multiple, complex statistical analyses, whose mechanisms are beyond the scope of this guide. The studies involved are then assessed individually for risk of bias, and the results are occasionally presented in a visual diagram similar to the example provided below. In the figure, seven different types of bias were assessed and listed on the y axis of the chart. All of the studies included in the review were then categorized into “low risk of bias, unclear risk of bias, or high risk of bias”. The percentage of studies that fall into each category are then plotted on the x axis for easy visual identification. Not all studies provide all of the information required to assess for bias, and not all NMAs report their bias analysis in graphical form. However, just as in traditional meta-analysis, it is imperative to read and understand the potential for bias included within the NMA, as



low-quality studies or high-risk bias risk studies can throw the conclusions into doubt and potentially prevent the clinician from drawing meaningful support from the study.

Tricco 2015

Once the calculations are completed, however, the question becomes “Can we trust the conclusions the calculations present?” Due to the fact that the conclusions reached by NMAs are the result of statistical calculations, rather than directly observed findings in a trial, the conclusions drawn must be felt to be valid by the clinician.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

The validity of network meta-analyses relies on three core principles: *homogeneity, similarity, and consistency*. Homogeneity refers to the analysis of the treatment effect of a single intervention within the meta-analysis. Put another way, homogeneity seeks to answer the question: “Does Drug A have roughly the same effect in each trial where it is studied?” If, on examination of ten trials of Drug A versus placebo, 5 trials found no benefit, and 5 trials showed benefit, those ten trials would be deemed heterogeneous, and data gleaned from their analysis would be suspect. However, if those ten trials all demonstrated a similar benefit, the data would be homogenous, and would be valid for inclusion in an NMA. The homogeneity of each intervention in the trial is assessed independently, meaning that if there are five interventions for comparison, the data homogeneity for each intervention would be assessed completely independently of one another. Homogeneity can be measured via multiple methods, including the use of a forest plot. If the majority of (and preferably all), studies have treatment effects that fall on the “benefit” half of the forest plot, the treatment effect is homogenous.

The second principle critical for determination of validity in an NMA is similarity. The principle of similarity is applied to all of the trials included in the NMA as a whole, and it seeks to assess how similar the study designs, disease severity, and base populations between the trials are. This makes sense conceptually as an NMA is essentially a gigantic hypothetical RCT where there are X number of treatment arms comparing different interventions against one another. If the base characteristics of the different study arms were significantly different than one another, there would be considerable risk of bias, and the conclusions of the RCT would be suspect. By the same token, if the base characteristics of the different studies are significantly different, then it is difficult to draw a valid comparison between the different interventions within the NMA. As with homogeneity, similarity can be measured with various tools or methodologies, with a common method being i^2 (i^2 represents the percentage variation between studies that is due to dissimilarity between studies rather than random chance [Higgins et al, 2003]). A NMA that has a high i^2 value likely has less valid conclusions than another study with a small i^2 value.

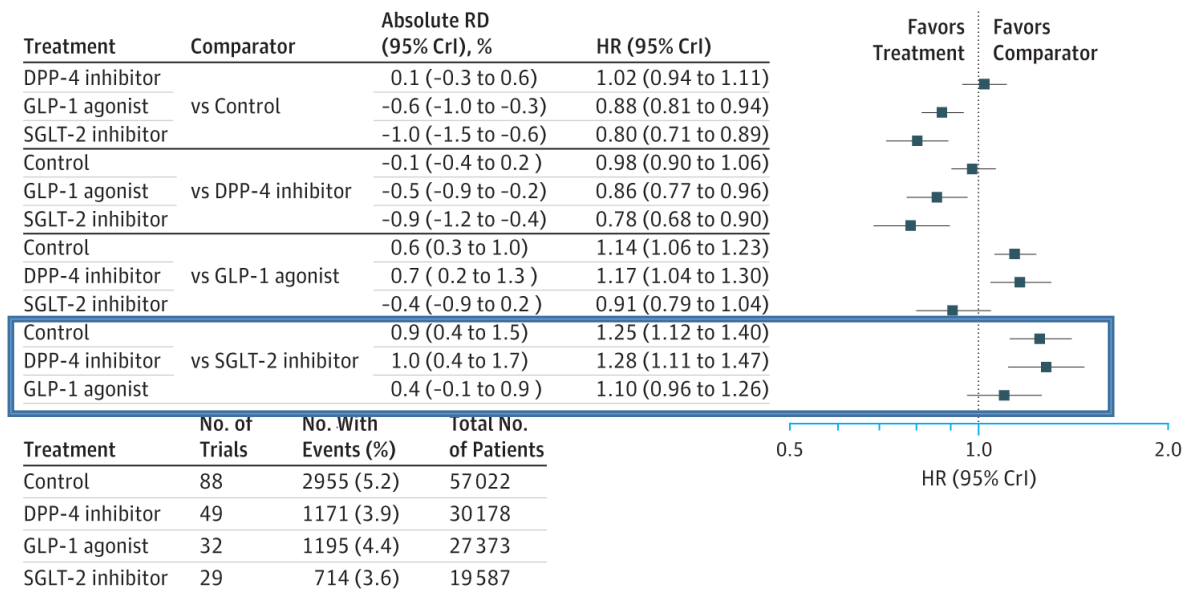
The third factor that helps ensure validity of a NMA is consistency. Often NMAs are generated because there is insufficient or poor-quality data comparing one or more interventions for a particular condition. However, there are often at least a few trials that will directly compare one intervention to another, and these studies can be used to ensure that the conclusions derived from the NMA are consistent with what has been directly observed in past trials. For example, if there were two trials comparing Drug A to Drug B showing no benefit for either drug, however, the NMA showed a strong benefit for Drug A compared to Drug B, the findings would not be consistent, and it would throw the conclusions of the NMA into doubt. This concept begs the question “If small, possibly poor-quality studies can overrule/cast doubt on the findings of a large NMA, what is the point of doing an NMA in the first place?” NMAs, when used correctly, can help support findings that may have already been found in smaller studies or may bring to light results that previously had not been studied. As with all research, conflicting studies prompt further questions and a need for further study.

How do you interpret a Network Meta-analysis?

Now that you (hopefully) have at least a conceptual understanding of what NMAs are, you may ask yourself: “How are NMA results presented, and how do I interpret those data?” Just as with traditional meta-analyses, there are a multitude of ways to conduct NMAs, and subsequently a multitude of ways that the data can be presented. In this section, we’ll attempt to explain two common presentations of NMA data, the Forest plot, and the League chart.

In the example below, you can see one example of how a study might illustrate its findings in a Forest plot. On the diagram, the primary outcome of “all-cause mortality” has been evaluated for three different therapies, as well as placebo. Each therapy has been compared individually to the other therapies available, and the hazard ratio results have been plotted in groups of three on the Forest plot to the right. In the section outlined in the blue box, we can see that SGLT-2 inhibitors have been compared to DPP-4 inhibitors as well as GLP-1 antagonists. The Forest plot clearly and easily demonstrates that SGLT-2 inhibitors were found to be superior to placebo as well as DPP-4 inhibitors but were not significantly better than GLP1 antagonists with respect to all-cause mortality.

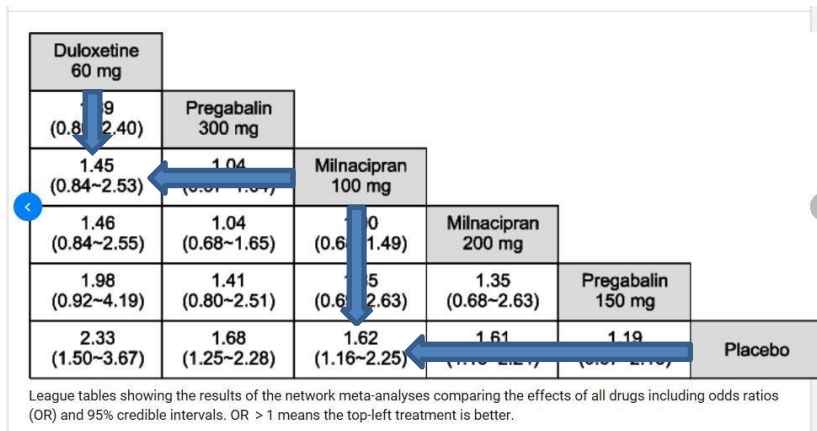
A Primary outcome: all-cause mortality, 97 trials; $I^2 = 12\%$



Zheng 2018

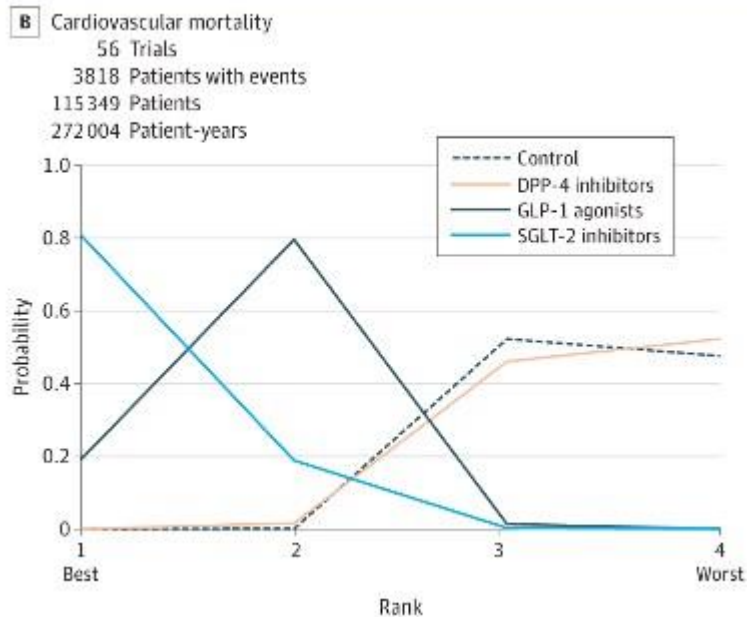
Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

The figure below is called a league chart and is a very common format for presenting data in NMAs. The columns and rows represent the different treatments being compared within the NMA. At the intersection of the column and rows, the comparative efficacy of the two treatments is reported as an odds ratio. The exact layout varies from study to study, but each table should include an explanation (as the chart does below) of which treatment is better. In the study below Duloxetine 60mg was not significantly better than Milnacipran 100mg as evidenced by the confidence intervals, but Milnacipran 100mg was better than placebo. Studies will often report separate League tables for each end point or indication being tested.



Young 2016

As the above examples demonstrate, the way data are reported in NMAs can lead naturally into attempting to “rank” various therapies. NMAs often will report “rankograms” that can (through statistics) try to determine which comparative therapy “ranks” better than others. These rankings are controversial as they report only a “probability” of one therapy being “better” than another, and it is often difficult to prove or replicate these findings. Rankograms and rankings from NMAs are NOT designed to provide definitive rankings or dictate appropriate treatment for clinicians. Ultimately the unique circumstances of each patient, as well as each patient’s tolerance will dictate what therapy is chosen. Rankograms can help suggest therapies, however, and can provide additional information if a clinician is attempting to decide between two equivalent therapies. For example, in the rankogram below, the NMA conducted suggests that SGLT-2 inhibitors have the highest likelihood of being the best therapy *of those tested* to prevent cardiovascular mortality.



Zheng 2018

Rankograms will usually be provided for each of the outcomes studied within the NMA and will typically reflect the results reported in the League charts or Forest plots. Therefore, it is important to take into account both the “raw” data reported in the head-to-head comparisons, as well as in the rankograms, as they may not always line up appropriately, and/or they may differ based on outcome studied. For example, Drug A may have the highest likelihood of being the best drug for the primary outcome, but Drug B may have the highest likelihood of avoiding adverse events. Therefore, it is important to take into account the needs and circumstances of your patient when making clinical therapy decisions.

Why should clinicians utilize Network Meta-analyses? What are their advantages?

Network meta-analysis is a unique and powerful tool for clinicians that goes beyond traditional meta-analysis and provides an additional support in the difficult challenge of clinical decision making. NMAs allow indirect comparison of multiple therapies, often at the same time, and can provide a “ranking” probability for the different therapies in order to help decide which therapies should be used first most often. NMAs can also help reduce the size of confidence intervals established in other studies. For example, if Drug A and Drug B were compared head-to-head in a one or two small studies, there may or may not be a benefit shown, and the confidence intervals may be quite broad. Incorporating that data into a network meta-analysis (that includes indirect comparison utilizing other common comparator studies) can allow for confirmation of an effect or non-effect, and/or can reduce the size of the confidence intervals by broadening the data pool.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

It should be noted, however, that the data and conclusions provided do not represent randomized data, as the subjects were not *actively* randomized by the NMA. The data remains randomized from the original studies, but an NMA cannot be considered a randomized trial. Therefore, conclusions drawn should be considered to be observational in nature, and a NMA cannot take the place of a large, prospective randomized trial. Despite this, NMAs can provide meaningful and insightful data that can assist clinicians, prompt further investigations, or demonstrate effects or connections not previously understood. As with all aspects of medicine, no one tool can be used alone to make decisions, but Network Meta-analysis can help improve patient care and help clinicians make the right decisions for their patients.

References:

1. Lee, Young. (2016). Overview of Network Meta-analysis for a Rheumatologist. *Journal of Rheumatic Diseases*. 23. 4. 10.4078/jrd.2016.23.1.4.
2. Zheng SL, Roddick AJ, Aghar-Jaffar R, Shun-Shin MJ, Francis D, Oliver N, Meeran K. Association Between Use of Sodium-Glucose Cotransporter 2 Inhibitors, Glucagon-like Peptide 1 Agonists, and Dipeptidyl Peptidase 4 Inhibitors With All-Cause Mortality in Patients With Type 2 Diabetes: A Systematic Review and Meta-analysis. *JAMA*. 2018 Apr 17;319(15):1580-1591. doi: 10.1001/jama.2018.3024. Review. PubMed PMID: 29677303
3. Tricco AC, Striffler L, Veroniki AA, Yazdi F, Khan PA, Scott A, Ng C, Antony J, Mrklas K, D'Souza J, Cardoso R, Straus SE. Comparative safety and effectiveness of long-acting inhaled agents for treating chronic obstructive pulmonary disease: a systematic review and network meta-analysis. *BMJ Open*. 2015 Oct 26;5(10):e009183. doi: 10.1136/bmjopen-2015-009183. Review. PubMed PMID: 26503392; PubMed Central PMCID: PMC4636655.
4. Riley RD, Jackson D, Salanti G, Burke DL, Price M, Kirkham J, White IR. [Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples](https://www.ncbi.nlm.nih.gov/pubmed/28903924). *BMJ*. 2017 Sep 13;358:j3932. doi: 10.1136/bmj.j3932. PubMed PMID: 28903924; PubMed Central PMCID: PMC5596393.

Submitted Jan 2019

II.8 Designing a study- comparing superiority and non-inferiority studies (Charlie Calliff, GSM4)

Background

When a researcher sets out to answer a specific question, one of the most important steps to doing so is selecting the correct study design. To determine which study design best fits the research question one is attempting to answer, the pros and cons of the possible study designs should enable one to make an educated decision. Randomized control trials are the main type of study if a researcher is interested in exploring the effect of a new intervention or treatment compared to a control or the current standard of care. The two main types of randomized control trials are superiority and non-inferiority trials. It can be challenging to know when to use one of these study designs versus the other, so here we will discuss to pros and cons of each and why a researcher may choose one over the other. For a more in-depth discussion of non-superiority trials, and some information on superiority trials, please look at section II.7 of the EBM guide.

Superiority trials

Superiority trials should be chosen when the goal of the project is to identify if an intervention is statistically significantly better than the standard of treatment or a control. *The outcome of a superiority trial is thus to establish efficacy*, with one possibility of causing changes to a practice or, if compelling enough, to practice guidelines. The null hypothesis in a superiority trial is that there is no difference between the new intervention and the control. When thinking about the study population of a superiority trial, superiority trials often require a smaller cohort to reach required power. Finally, in terms of understanding the results of a study, superiority trials are presented based on p-values, typically with confidence intervals, making it simple to understand if a result is statistically significant. The data also lends itself to reporting results as relative or absolute risk reduction (or increase), and the derivative of NNT (number needed to treat).

In terms of challenges or drawbacks of superiority trials, there are two main drawbacks. One is the ethical dilemma involved in randomizing patients who may require care to a control or inferior treatment group. There are safeguards in place to protect against some of the harm associated with this randomization, such as ending trials early if efficacy has been established prior to reaching the designated study endpoint. The second drawback is, it is possible that companies or researchers may invest a lot of resources towards developing a new trial. If the trial does not find a statistically significant difference, although it provides valuable data, it can be a major loss of capital and time. Additionally, significant bias can be introduced when a drug is being investigated when a researcher or industry is not neutral.

Non-inferiority trials

Non-inferiority trials should be chosen when the goal of the project is to identify if an intervention is not statistically significantly worse than the standard of care currently in practice. Instead of looking at efficacy, one is looking at the similarity of treatments. *The null hypothesis for a non-inferiority trial is that the new intervention is not equivalent to the standard of care.* This study design may be desirable if a new treatment has some added benefit besides efficacy, such as fewer side effects, increased safety, different routes of delivery, etc. Non-inferiority trials can be regarded as less ethically ambiguous because patients in the control group do not go without treatment.

In terms of challenges or drawbacks associated with non-inferiority trials, one such challenge is that they can be difficult to interpret. Specifically, to determine non-inferiority, one must choose a non-inferiority margin which by nature has a degree of subjectivity (see section II.7). Some important factors for determining a non-inferiority margin include evaluating previous studies, clinical relevance, and practicality related to available sample size. Based on the agreed-upon non-inferiority margin, the result is determined to be inferior or non-inferior compared to standard care. Second, compared to superiority trials, non-inferiority trials typically require a larger sample size to reach power.

Overall, there are a few main aspects of both trial types that need to be weighed to choose between designing a superiority versus a non-inferiority trial. *The major difference between the two study types is the desired endpoint, evaluating efficacy or similarity.* Additionally, by understanding the drawbacks of each study type, one can better decide which design best meets their needs and the requirements of answering a specific research question. Unfortunately, no formula will tell you exactly which study design to use. Instead, the choice between these two study types largely comes down to the research question that is being addressed and the clinical background that it is being answered within.

Kishore and Mahajan, 2020, include a great table in their paper which summarizes many of the topics discussed above:

	Type of trial	
Characteristics	Superiority trial	Noninferiority trial
Condition for application	Comparing novel intervention with non-standardized intervention	Comparing novel intervention with standard intervention
Control	Placebo or non-standard Intervention	Standard intervention
Intervention	Novel intervention	Novel intervention
Hypothesis	Two-tailed	One-tailed
Key to hypothesis formation	Effect size (d)	Effect size (d) and clinically meaningful difference (Δ)
Statistical Significance range	$\mu_1 - \mu_0 \neq 0$	$\mu_1 - \mu_0 > -\Delta$
Analysis recommendation	ITT	ITT and per protocol analysis
Reporting	<i>P</i> -value and CI	CI

References

Christensen E. Methodology of superiority vs. equivalence trials and non-inferiority trials. J Hepatol. 2007 May;46(5):947-54.

Kishore K, Mahajan R. Understanding Superiority, Noninferiority, and Equivalence for Clinical Trials. Indian Dermatol Online J. 2020 Sep 19;11(6):890-894.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Wang B, Wang H, Tu XM, Feng C. Comparisons of Superiority, Non-inferiority, and Equivalence Trials. Shanghai Arch Psychiatry. 2017 Dec 25;29(6):385-388.

II.9 Non-Inferiority Trials (Lukas Emery)

What it is not

Superiority trials—what we are used to seeing. Study compares a treatment to either placebo or existing gold standard and shows a statistically significant superiority in the results

Equivalence trials—typically used to show there really is no significant difference between two versions of the same drug, e.g., generic drugs or vaccine lots.

What is a Non-inferiority Trial

Background:

Noninferiority trials are an important tool for the evaluation of many therapeutic interventions such as new drugs or biologics, medical devices, and a wide variety of other therapies. The trial design allows one to circumvent the standard placebo or no-treatment control as this is not ethical when many conditions already have an effective treatment established. Therefore, noninferiority trials seek to compare new interventions to existing therapies/standard of care in an effort to prove that their efficacy is “not inferior” to currently available treatments. The ultimate goal is to determine that a new intervention is not worse than a control treatment (i.e., some existing therapy) by a reasonably small amount with an acceptable degree of confidence.

Trial Design:

The null hypothesis in a noninferiority study states that the primary end point for the experimental treatment is worse than that for the control treatment by a prespecified margin (inferiority margin). Rejection of the null hypothesis would, therefore, support the claim that a new intervention is not inferior to the comparison therapy. The foundation of noninferiority trials is built on several factors:

1. **RCTs involving control:** The availability of randomized control trials showing superiority of the control treatment compared to placebo.
2. **Establishing Endpoints:** Researchers must select an appropriate endpoint to be studied; once this has been established available data is used to determine the expected performance/efficacy of the control treatment.
3. **Setting the Non-inferiority Margin:** A threshold below which it can be established that the new drug is not worse than its comparator. This is based on both statistical and clinical considerations as outlined below.

Determining the Noninferiority Margin:

This margin should be chosen such that the new drug can be considered to be effective relative to placebo (even when a placebo group is not included) and needs to account for the uncertainty in the effect size of the active control versus placebo. First, you need to make a “constancy assumption” (i.e., effect shown in prior studies will be consistent in your noninferiority study) about the effectiveness of the control compared to placebo as this will not be assessed in the non-inferiority trial; therefore, more data about comparator = more precise estimate of effect. In general, this is a conservative estimate of the effect of the comparator based on available data and usually represents the smallest effect size. Researchers must then, using clinical judgment, determine a clinically acceptable difference (degree of noninferiority) of the test drug compared to the active control (i.e., how much of the treatment effect needs to be preserved). This consideration is often related to the seriousness of the outcome, the benefit of the active comparator and the relative safety profiles of the test drug and the comparator. The higher the percentage to be preserved the more conservative the noninferiority margin, thus making it more difficult to conclude noninferiority. For more detailed information on setting the noninferiority margin please see excerpt below:

Excerpt From: Wangge G, et al.

Most of the guidelines on noninferiority trials state that a margin should account for both clinical and statistical considerations. However, details on how such a margin should be determined are not clearly specified, with the exception of the recently drafted guideline on noninferiority trials issued by the FDA. The guideline was composed based on previous guidelines and methodological publications on noninferiority trials published since the 1980s. The guideline is only one example of determining a noninferiority margin, and it reflects regulatory interest; thus, its focus is on showing indirect efficacy of the test drug compared with placebo.

The guideline recommends the fixed-margin method, or 95%–95% method, which is considered the most straightforward and readily understood approach. The method starts by identifying M1 and M2. M1 is the effect of the active control compared with placebo, which is assumed to be present in the noninferiority trial. M1 is chosen as a conservative estimate (smallest effect size possible) of the effect of the active comparator, which is the upper bound of the 95% confidence interval (CI) of the pooled effect size, rather than the point estimate. M2 reflects the clinical judgement about how much of M1 should be preserved and represents the largest clinically acceptable difference (degree of inferiority) of the test drug compared with the active control. For example, if it is necessary that a test drug preserve 75% of a mortality effect, M2 would be 25% of M1, the loss of effect that must be ruled out. Determining M2 assures that the test drug will be superior to placebo.

Determining M1, as the first step in defining a noninferiority margin, can be based on one or more placebo-controlled trials of the active comparator that have a design similar to the current noninferiority trial. A meta-analysis of several placebo-controlled trials is preferable, because it will result in a pooled, more precise effect estimate of the active comparator.

The second step is to calculate M2 from M1 by choosing a certain amount of the effect to be preserved. The draft FDA guideline implicitly recommends using a preserved effect of

50% to determine M2. Choosing a higher percentage to be preserved (e.g., 67%, where M2 is 33% of M1) results in a stricter or more conservative noninferiority margin, meaning it is more difficult to conclude noninferiority. The formula to calculate M2 for a risk difference (RD) is:

$$(1 - \text{preserved effects}) \times M1$$

For the relative risk (RR), and other ratio measures, the guideline discusses 3 methods for calculating M2. The preferred method calculates the margin using the natural logarithm:

$$e^{\ln(1/M1) \times (1 - \text{preserved effects})} \text{ or } (1/M1)^{(1 - \text{preserved effects})}$$

Interpreting the Results:

The results of the noninferiority trial are compared with the prespecified noninferiority margin as follows: if the upper bound of the 95% CI for the effect estimate is smaller than the noninferiority margin, noninferiority is concluded. For example, if a noninferiority trial shows that the RR of the new drug compared with the active comparator is 0.94 (95% CI 0.72 to 1.25), and the noninferiority margin is 1.3, it can be concluded that the new drug is noninferior to the active comparator.

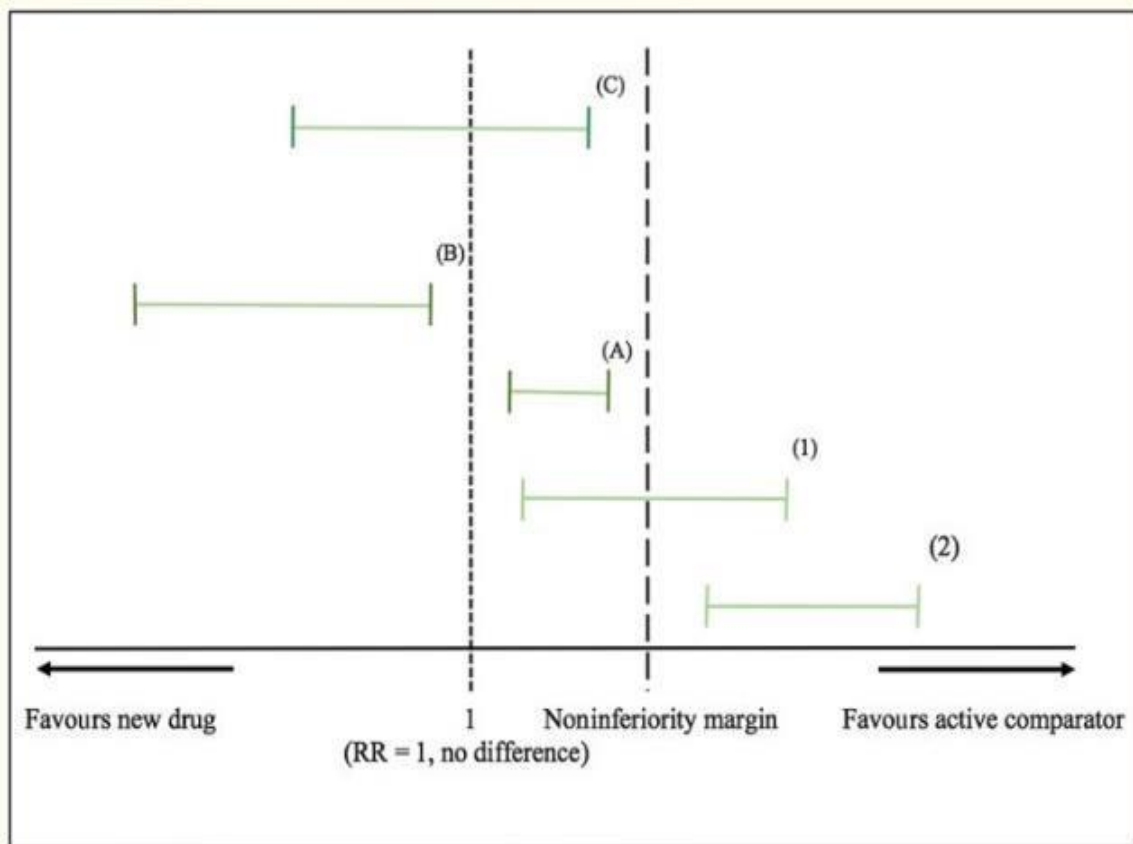


Figure 1

Analysing noninferiority by comparing the confidence interval (CI) of the relative risk to a predefined margin. (1) and (2) Noninferiority was not demonstrated because the upper limit of the CI exceeded the margin. (A), (B), (C) Noninferiority was demonstrated because the upper limits of the CI did not exceed the margin

Althunian TA, et al. Defining the noninferiority margin and analysing noninferiority: An overview. *Br J Clin Pharmacol.* 2017;83(8):1636-1642.

Issues/Limitations of noninferiority trials:

1. Lack of placebo group and reliance on “constancy assumptions” based on prior published data for comparator effect.
2. Variation in noninferiority margins chosen for the study.
 - One can easily see how setting a less conservative margin can lead to the finding of “noninferiority” when in fact, the results are just a reflection of a poorly prespecified noninferiority margin.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

3. Reliance on subjective factors (i.e., clinical judgement) when determining an appropriate preserved-effect value again influencing the noninferiority margin.
 - This is particularly challenging when using noninferiority design for safety studies as there are usually no reasonable data to justify the margin for safety; instead, the researchers must decide what level of adverse events is acceptable.

References:

1. L. Mauri, R.B. D'Agostino Sr. Challenges in the design and interpretation of noninferiority trials *N Engl J Med*, 377 (October (14)) (2017), pp. 1357-1367.
2. Althunian TA, de Boer A, Groenwold RHH, Klungel OH. Defining the noninferiority margin and analysing noninferiority: An overview. *Br J Clin Pharmacol*. 2017;83(8):1636-1642.
3. Wangge G, Roes KC, de Boer A, Hoes AW, Knol MJ. The challenges of determining noninferiority margins: a case study of noninferiority randomized controlled trials of novel oral anticoagulants. *CMAJ*. 2013;185(3):222-7.

Submitted November 2018

II.10 Pragmatic Clinical Trials- What Are They? (Priya Katari)

Key attributes of PCTs:

1. **intent to inform decision-makers** (patients, clinicians, administrators, and policymakers), rather than clarifying a biological or social mechanism
2. **an intent to enroll a population relevant to the decision in practice and representative of the patients/populations and clinical settings for whom the decision is relevant;** and
3. an intent to either
 - a. **streamline procedures and data collection** so that the trial can focus on adequate power for informing the clinical and policy decisions targeted by the trial **or**
 - b. **measure a broad range of outcomes.**

Common sense definition for a PCT would thus be as follows:

“Designed for the primary purpose of informing decision-makers regarding the comparative balance of benefits, burdens and risks of a biomedical or behavioral health intervention at the individual or population level.”

Submitted February 2019

Editor’s note: Dr. Katari introduced bookmarks and hyperlinks to this document, materially enhancing its usability.

Flexibility in adherence: The trial is flexible in how the users engage with the intervention and does not have special measures to enforce adherence.

Follow-up: The trial has no more follow-up than usual care and limits additional data collection.

Primary outcome: The trial has an outcome that has the most recognizable importance to the participants and measures the outcome in a way that is similar to usual care.

Primary analysis: The trial implements an intention-to-treat analysis using all available data.

In summary, when designing a clinical trial, closely mimicking what happens in usual care will lead to a higher PRECIS-2 score, or higher pragmatism. It is also important to note that pragmatic trials are not free of limitations, and very few trials are truly pragmatic on all nine domains [3]. Rather than categorizing trials as either explanatory or pragmatic, it is helpful to view pragmatism as a continuum, as the PRECIS-2 tool illustrates. Furthermore, researchers should not be discouraged from designing trials that lean toward the explanatory end. Rather, researchers should design trials that fit their intended purpose.

References:

1. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol.* 2009;62(5):464-475. doi:10.1016/j.jclinepi.2008.12.011
2. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ.* 2015;350:h2147. Published 2015 May 8. doi:10.1136/bmj.h2147
3. Ford I, Norrie J. Pragmatic Trials. *N Engl J Med.* 2016;375(5):454-463. doi:10.1056/NEJMra1510059

Submitted 2-2022

II.12 Phases of New Drug Investigation Trials– (Katie Kozacka, GSM4)

If you're trying to nail down the differences and characteristics of various trials in study, the following should help!!

A. Phase 0: "Exploring If and How a New Drug Works."¹

- This type of study is not commonly used.
- A few small doses are used on a few individuals who likely do not benefit from this treatment.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

- Instead, the purpose is to speed up approval processing and to help others in the future.
- Generally, this type of study looks more at how a drug reacts with a target organ, tissue, or how it is distributed in the body.
- Sometimes, this could require a biopsy, sample, or testing of the participant to evaluate these interactions.
- This is not a required part of testing for drug approval.

B. Phase I: "Is the Treatment Safe?" "First in Human Studies."²

- The goal of this phase is to determine a suitable dose for phase II and to test safety of the drug.
- Minimal dose for toxicity and maximum tolerated dose are defined.
- Even if the drug has already undergone animal testing, effects and distribution may be different in human studies.
- A small dose is given to a few patients to start. Then as tolerated, dosing increases by 100%, 66%, 50%, 40%, 33% etc. until severe or dose limiting toxicity in a large fraction of the participants ends the trial.
- Many subjects will in the end receive sub-therapeutic dosing and will not be able to have benefits from the drug.
- Titration may not occur in one single participant because then the effects of dosing cannot be distinguished from long term side effects of the drug. The phase I trials are not good at picking up time dependent side effects or rare toxicities.

C. Phase II: "Does the Treatment Work?"¹²

- IIA: Treatment is given to a small group of patients 12-100 at one strong dose.
- IIB: Treatment is given in several doses to assess optimal dose.
- Phase II involves a much larger group of patients.
- Less common side effects can be picked up in this way.
- No placebo is used.
- If enough benefit from treatment, the drug moves on to phase III.

D. Phase III: "Is It Better than What is Available?"¹²

- Last testing before being submitted to the FDA for approval.
- Large number of patients, longer duration, greater scope.
- Placebo or standard of care used.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

- This study can confirm dosing, timing, and frequency. It is used for the package insert/drug leaflet.
- Confident efficacy evaluation. Also finds more toxicities.
- If passes stage III, a New Drug Application form is submitted for approval.

E. Phase IV: “What Else Do We Need to Know?”¹

- Used for drugs already FDA approved and is therefore the safest type of study.
- Looks at other aspects of the treatment such as quality of life or cost.

References:

1. “What are the Phases of Clinical Trials?” *American Cancer Society*. February 2017.
<https://www.cancer.org/treatment/treatments-and-side-effects/clinical-trials/what-you-need-to-know/phases-of-clinical-trials.html>
2. Brody, Tom, “Clinical Trial Design,” *Clinical Trials Second Edition*, 2016 Elsevier, Chapter 2, 31-68.

Submitted January 2019

II.13 Understanding Endpoints with an emphasis on cancer trials (David Lakomy)

Q: What is the key requirement for new cancer drug approval?

A: Basically, the end goal is to demonstrate *efficacy* with acceptable *safety*.

Q: But I have read plenty of studies that have used a variety of endpoints that didn't directly test efficacy?

A: Cancer drug trials go through several phases (please see "Phases of New Drug Investigation Trials" for more detailed information) prior to approval. In brief, phase I trials evaluate toxicity and tolerability, phase II trials determine anti-tumor activity, and phase III determine clinical benefit. Thus, different stages of clinical trials require different endpoints with early phase trials testing for endpoints regarding pharmacokinetics, pharmacodynamics, and tumor shrinkage and later stage trials testing for patient centered efficacy in terms of prolongation of survival or improvement in symptoms.

Q: That is confusing, lets break it down further step-by-step, what endpoints are there for phase I trials?

A: The conventional primary endpoints of phase 1 trials have historically been: maximum tolerated dose (MTD), recommended phase 2 dose (RP2D), and estimation of safety profile of the new drug.

The MTD is determined by the occurrence of dose-limiting toxicities (DLTs) defined by the occurrence of severe toxicities during the first cycle of systemic cancer therapy. Such toxicities are assessed according to the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) classification, and usually encompass all grade 3 or higher toxicities with the exception of grade 3 nonfebrile neutropenia and alopecia.

The RP2D then, is usually the highest dose with acceptable toxicity, usually defined as the dose level producing around 20% of dose-limiting toxicity.

Q: That seems fairly straightforward, are there any issues with using MTD, DLT, and RP2D in phase I trials?

A: There are several. For one, the DLT definition stated above, while still the most commonly used, is met with a fair degree of heterogeneity in terms of its criteria and how it is applied in patient studies. There is no singular consensus on the definition of DLT in phase I trials.

Secondly, and more profoundly, this standard is largely based on cytotoxic chemotherapy drugs and regimens that dominated oncology for decades but are now less applicable in our targeted molecular therapy age. For example, chemotherapies were administered for a set period of time (in cycles) as opposed to often continuously for novel molecular therapies. In turn, some lower grade (\leq grade 2) toxicities that may have been passable if they were experienced only transiently may become intolerable if they are experienced continuously (e.g., long-term low-grade diarrhea or xerostomia).

RP2D may also be affected by this history tied to chemotherapy drugs. While there is typically a direct relationship between dose and efficacy for chemotherapeutic agents (i.e., higher dose resulting in greater efficacy), for molecular agents this is not always the case and lower doses with similar efficacy may produce lower toxicity.

Overall, this remains an evolving field.

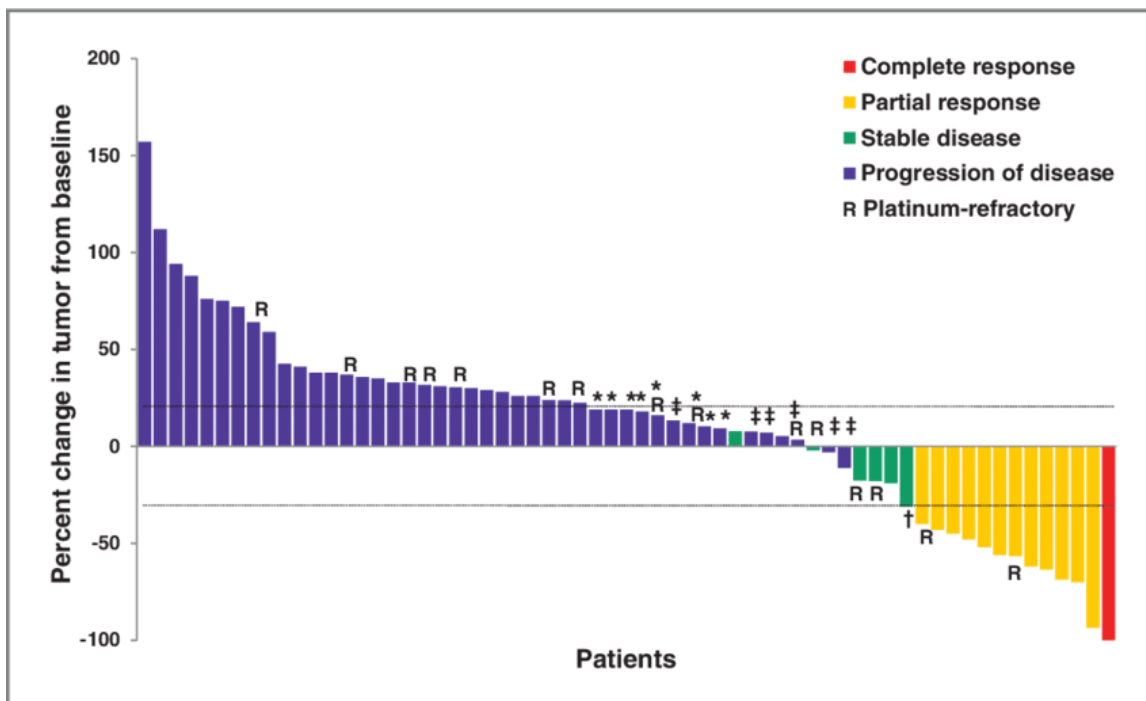
Q: Okay, so what about phase II trials, what are the endpoints here?

A: Phase II trials begin to answer the question of whether or not the drug will work, that is for oncology trials, does this drug have anti-tumor activity in humans. Thus, tumor response measured as objective response rate (ORR) or progression-free survival (PFS).

Q: How is ORR determined and analyzed?

A: ORR is defined as the proportion of patients with tumor size reduction of a predefined amount and for a minimum time period. Response duration usually is measured from the time of initial response until documented tumor progression.

While a variety of criteria exist, for solid tumors the Response evaluation criteria in solid tumors (RECIST) guidelines are the most commonly applied. RECISTS consists of identification and classification of tumor lesions, periodic assessment (usually radiographic), comparison to baseline, and placement of tumor response into different categories: complete response (CR), partial response (PR), stable disease (SD),



progressive disease (PD), and not evaluable (NE).

One way to depict this and to demonstrate maximal changes in tumor size is a waterfall plot (see right) with each patient representing a column and the magnitude of change organized by magnitude of change.

Q: What about PFS?

A: PFS is defined as the time from randomization until objective tumor progression or death, whichever occurs first.

The issue with both ORR and PFS is that “progression” and “response” are difficult to standardize both in the fact that “disease progression” can be collected from multiple sources (including physical exams at unscheduled visits and radiological scans of various types) and at different times, and that both physical exam and radiographic interpretation are susceptible to subjective errors.

Phase III trials are best at determining patient safety and efficacy.

Q: So, if phase III trials are the best for determining patient safety and efficacy, what are the most appropriate endpoints?

A: Everything I talked about above (ORR, PFS) are tumor-centered endpoints, that is they are centered on how the tumor reacts to treatment, for phase III trials what we aim to look at is patient-centered endpoints: overall survival (OS) and quality of life (QoL).

The primary goal of cancer treatment is to provide a cure and prolong life, thus, OS remains the gold-standard for demonstrating clinical benefit. OS is defined as the time from randomization until death from any cause and is measured in the intent-to-treat population. Survival is the most reliable cancer endpoint and is not subject to bias (if patients are appropriately evaluated in randomized controlled studies with similar baseline characteristics).

QoL constitutes the other patient-centered oncologic endpoint. QoL is any report of the status of the patient’s health that comes directly from the patient, without interpretation of the patient’s response by a clinician or anyone else and may include symptoms, functioning, or a more global assessment of the effect of the disease on health and overall functionality of the patient. Most often this is reported through the use of validated survey instruments. While there is growing emphasis on the evaluation of QoL, additional standardization and utilization is needed across clinical trials.

Q: Hold-up, I get why OS would be considered the gold-standard for phase III trials, but I have read a ton of papers with other endpoints, what gives?

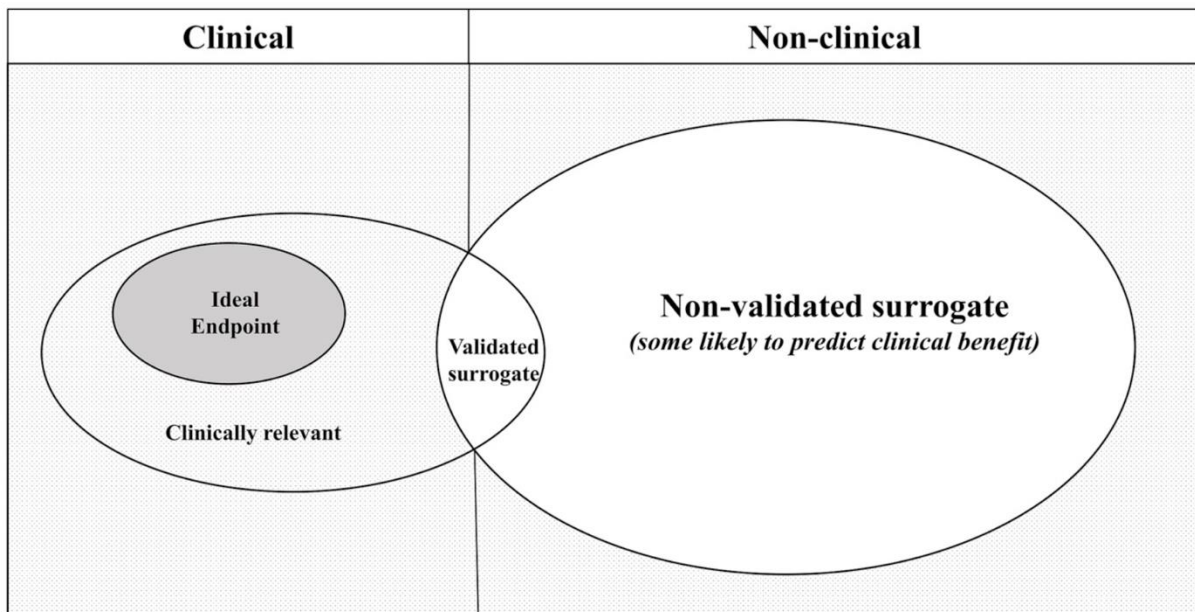
A: This gets to the crux of a complex and messy topic really quickly. While this deserves its own fully fleshed out exploration, I will quickly discuss it here. Basically, the goal of phase III trials is to generate evidence that can guide clinical decision making for patients, physicians, and policy makers. There are nearly endless outcomes that could be measured, and a single endpoint (even OS) does not provide a fully fleshed out picture from which to make all decisions.

Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine

As discussed above, endpoints can largely be split into two: patient-centered (associated with how a patient feels, functions, and survives) and tumor-centered (pathologic response, biomarker change, ORR, PFS, etc.). As this later group does not directly measure patient-derived outcomes, they may be considered *surrogate endpoints*. Surrogate endpoints are closely associated with clinically meaningful endpoints and thus taken to be a reliable substitute for them. This raises the question as to why anyone would use these endpoints at all. In short, as oncologic treatment and follow-up can often have a very prolonged course, assessing OS may prove difficult if not impossible.

In an ideal world, more-easily accessible surrogate endpoints would all serve as a proxy for more clinically meaningful endpoints, but this is not the case. The Prentice criteria have been developed to test for validity with the criteria as followed: the treatment has an effect on survival time, the treatment has an effect on the surrogate, the surrogate is associated with survival time, and the treatment effect on survival is captured by the surrogate. A list of validated surrogates for both oncologic and other clinical trials can be found here: <https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure>

This can get complicated very quickly and has different nuances and specifics for different cancer sites, but in sum while clinically relevant, personally meaningful endpoints are ideal, endpoints remain on a spectrum. Below is a figure depicting these variations.



- Meaningful endpoint
- Potentially meaningful endpoint
- Meaningless endpoint

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Q: So how do I interpret PFS vs OS vs ORR vs all the other endpoints I have seen used in phase II/III trials?

A: Here is a chart to help compare some of the most common endpoints you will come across in later oncologic trials:

Endpoint	Definition	Advantages	Disadvantages
Overall Survival	Time from randomization until death from any cause	<ul style="list-style-type: none"> • Easily and precisely measured • Generally based on objective and quantitative assessment 	<ul style="list-style-type: none"> • May be affected by switch-over of control to treatment or subsequent therapies • Needs longer follow-up • Includes noncancer deaths
Disease-free survival (and event-free survival) [DFS]	Time from randomization until disease recurrence or death from any cause	<ul style="list-style-type: none"> • Generally assessed earlier and with smaller sample size compared with survival studies • Generally based on objective and quantitative assessment 	<ul style="list-style-type: none"> • Potentially subject to assessment bias, particularly in open-label studies • Definitions vary among studies • Balanced timing of assessments among treatment arms is critical • Includes noncancer deaths
Objective response rate [ORR]	Proportion of patients with tumor size reduction of a predefined amount and for a minimum time period	<ul style="list-style-type: none"> • Generally assessed earlier and with smaller sample size compared with survival studies • Effect on tumor attributable to drug(s), not natural history • Generally based on objective and quantitative assessment 	<ul style="list-style-type: none"> • Definitions vary among studies • Frequent radiological or other assessments • May not always correlate with survival
Complete response [CR]	No detectable evidence of tumor	<ul style="list-style-type: none"> • Generally assessed earlier and with smaller sample size compared with survival studies • Effect on tumor attributable to drug(s), not natural history • Generally based on objective and quantitative assessment 	<ul style="list-style-type: none"> • Definitions vary among studies • Frequent radiological or other assessments • May not always correlate with survival
Progression-free survival [PFS]	Time from randomization until objective tumor progression or death, whichever occurs first	<ul style="list-style-type: none"> • Generally assessed earlier and with smaller sample size compared with survival studies • Measurement of stable disease included • Generally based on objective and quantitative assessment 	<ul style="list-style-type: none"> • Potentially subject to assessment bias, particularly in open-label studies • Definitions vary among studies • Frequent radiological or other assessments • Balanced timing of assessments among treatment arms is critical • May not always correlate with survival

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Symptomatic symptom endpoints	No singular definition but generally patient symptom assessments and/or physical signs representing symptomatic improvement	<ul style="list-style-type: none"> • Generally assessed earlier and with smaller sample size compared with survival studies 	<ul style="list-style-type: none"> • Blinding is important for assessing the endpoint • Potentially subject to assessment bias, particularly in open-label studies • Lack of validated instruments in many disease areas • Definitions vary among studies • Balanced timing of assessments among treatment arms is critical
-------------------------------	---	--	--

Q: So, in the end, how am I supposed to interpret papers in relationship to patients?

A: There are no easy answers. Optimization and selection of endpoints for clinical trials is an evolving field, especially with the onset of novel molecular therapies. It is prudent that every physician has an understanding of the range of endpoints available, the context in which they have arisen, their strengths and limitations, and how those intersect with the clinical specifics and personal values of each patient.

Q: Where could I find more details about these topics?

A: References:

1. Anagnostou V, Yarchoan M, Hansen AR, et al. Immuno-oncology Trial Endpoints: Capturing Clinically Meaningful Activity. *Clin Cancer Res*. 2017;23(17):4959-4969. doi:10.1158/1078-0432.CCR-16-3065
2. Kilickap S, Demirci U, Karadurmus N, Dogan M, Akinci B, Sendur MAN. Endpoints in oncology clinical trials. *J BUON*. 2018;23(7):1-6.
3. McLeod C, Norman R, Litton E, Saville BR, Webb S, Snelling TL. Choosing primary endpoints for clinical trials of health care interventions. *Contemp Clin Trials Commun*. 2019;16:100486. Published 2019 Nov 12. doi:10.1016/j.conctc.2019.100486
4. Postel-Vinay S. Redefining dose-limiting toxicity. *Clin Adv Hematol Oncol*. 2015;13(2):87-89.
5. Tannock I, Aaamdal S, Arnold D, et al., Clinical Trial Endpoints. European Society for Medical Oncology: Educational Portal for Oncologist. 2015 <https://oncologypro.esmo.org/education-library/clinical-trial-resources/tips-and-tricks>
6. U.S. Department of Health and Human Services, Food and Drug Administration, Oncology Center of Excellence. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics: Guidance for Industry. 2018. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics>

Submitted 12/28/2020

Section III. Fundamental Research Methods and Statistics

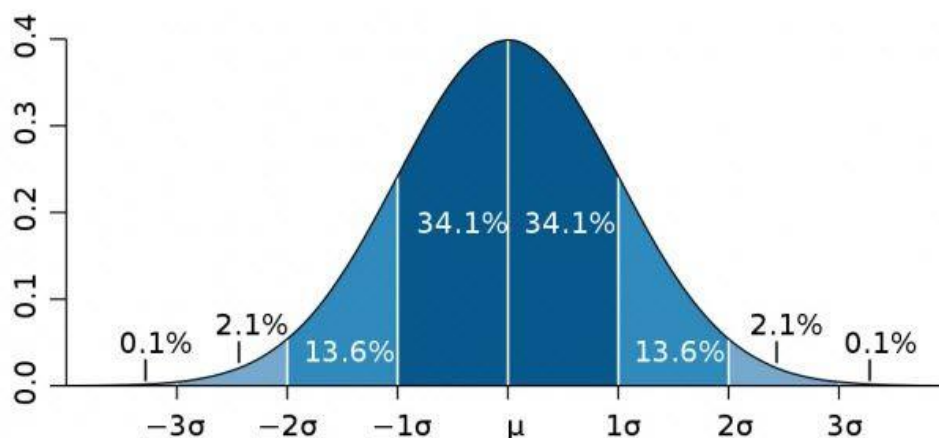
III.1 The Bell Curve – What is a “normal distribution” and why does it matter? (Chad Y. Lewis, GSM4)

When discussing datasets, we often hear the term “normal” (or “Gaussian”) distribution being thrown around but may not fully understand what a true “Bell Curve” is. More importantly, when designing a study or interpreting the validity of someone else’s study, it is helpful to understand how data distributions affect which tools should be used for statistical analyses.

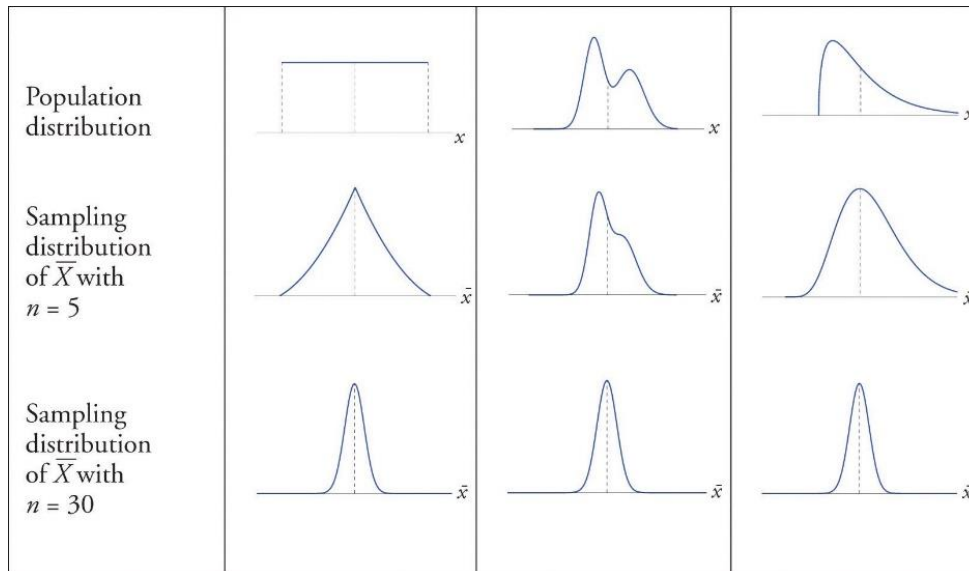
A “standard normal distribution, or “Bell curve”, has a mean, median, and mode equal to zero, the curve is symmetric at the center, and the standard deviation is one. This creates predictable areas under the curve (adding up to one) along the X-axis which can be used to quantify percentages of a dataset or determine probability of events (i.e., z-value) (Figure 1).¹ Many medicine-related variables such as blood pressure or HbA1c will naturally follow a normal distribution, however this is not always the case. For example, a variable such as bacterial growth exhibits an exponential curve and would not fit a normal distribution. There are many other data types that naturally follow a non-normal distribution.² Examples include:

- Weibull distributions – found with data such as average survival time given a diagnosis.
- Log-normal distributions – found with length data such as height.
- Poisson distributions – found with rare events such as number of accidents.
- Binomial distributions – found with “proportion” data such as percent of birth defects.

Figure 1 – The Bell Curve



There are statistical tools such as Chi-square (and less commonly, Kolmogorov-Smirnov or Shapiro-Wilk)



that can measure the normality or skew of a dataset.³ However, statisticians often rely on a rule of thumb called the “central limit theorem (CLT).” The CLT postulates that a sample size (n) of 30 or greater will approximate a normal distribution, even if the population distribution being studied is not normal (Figures 2 and 3).^{4,5} Therefore, you should keep in mind that even outside of Texas, “bigger is better” when it comes to sample size!

Figure 2 – Demonstration of CLT with an n of 5 vs. 30

Some commonly used statistical tools such as t-tests and analysis of variance (ANOVA) require a normal distribution of data to function appropriately. When data is not normally distributed, the cause should be determined, and corrective actions should be taken if applicable.²

While not within the scope of this chapter, there are advanced statistical tools that can “transform” or normalize a dataset to make it fit a normal distribution. Here are some common causes of non-normality and their respective corrective actions:

- **Extreme Values:** Too many extreme values in a data set will skew the data distribution. This can be rectified by “cleaning” the data by determining measurement errors, data-entry errors, and outliers and removing them from the data (for valid reasons).
- **Overlap of Two or More Processes:** This can occur when data comes from more than one process or from a process that changes frequently. If two or more data sets that would be normally distributed on their own are overlapped, data may look bimodal or multimodal (two or more most-frequent values). In these situations, determine which X’s cause the bimodal or multimodal distribution and then stratify the data.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

- **Insufficient Data Discrimination:** Rounding errors or imprecise measurements can make truly continuous and normally distributed data look discrete and non-normal. This can be overcome by using more accurate measurement systems or by collecting more data.
- **Values Close to Zero or a Natural Limit:** If a process has many values close to zero or a natural limit, the data distribution will skew to the right or left.

Using an example based on one of the above-listed causes of non-normality (extreme values), let us consider how this could play out in a real-world situation. Say that we wanted to look at the “average” wealth of American households. One can extrapolate that given the highly disparate concentration of wealth within the top 5% of Americans in comparison to the other 95%, this would create a significantly skewed distribution of data and may not give us a practical or useful average if we were to look at group means.

In this example, it would be more appropriate to look at group medians to minimize the impact of extreme outliers. This is where “non-parametric” tests come into play. It is a bit of an oversimplification, but generally one can imagine that parametric tests should be used to test group means and non-parametric tests should be used to test group medians.⁶ Since non-parametric tests do not require continuous data (as parametric tests typically do), they can also be used to analyze variables such as ordinal or ranked data. Whichever parametric test you are used to seeing, you can bet that there is likely an equivalent non-parametric statistical tool that can be used for non-normal data (Table 1).

Table 1 - Comparison of Statistical Analysis Tools for Normal vs. Non-Normal Distributions

Tools for Normally Distributed Data (Parametric)	Equivalent Tools for Non-Normally Distributed Data (Non-parametric)
T-test	Mann-Whitney test; Mood’s median test; Kruskal-Wallis test
Paired t-test	One-sample sign test
ANOVA	Mood’s median test; Kruskal-Wallis test

In summary, the utility of statistics is based on the impossibility of collecting data from an entire population. Rather, by taking a sample of data from a subset of the larger population, we can then extrapolate and draw conclusions about the population. Because of this, practices such as standard hypothesis testing often assume that the population data is normally distributed. Therefore, one must be aware of the normality of the dataset they are working with (i.e., “goodness of fit” to a Bell curve), the limitations of the type of hypothesis test being used (i.e., parametric vs. non-parametric), or at a minimum have a sample size sufficiently large enough to rely on the Central Limit Theorem when seeking to conduct a valid study.⁵

References:

1. Glen S. Normal Distributions (Bell Curve): Definition, Word Problems. StatisticsHowTo.com. <https://www.statisticshowto.com/probability-and-statistics/normal-distributions/>. Accessed January 2, 2021.

2. Buthmann A. DEALING WITH NON-NORMAL DATA: STRATEGIES AND TOOLS. ISixSigma. <https://www.isixsigma.com/tools-templates/normality/dealing-non-normal-data-strategies-and-tools/>.
3. Glen S. Goodness of Fit Test: What is it? StatisticsHowTo.com. <https://www.statisticshowto.com/goodness-of-fit-test/>. Accessed February 2, 2021.
4. Frost J. Central Limit Theorem Explained. Statistics By Jim. <https://statisticsbyjim.com/basics/central-limit-theorem/#:~:text=The central limit theorem in,variable's distribution in the population.>
5. Singh S. Central Limit Theorem Simplified! <https://medium.com/@seema.singh/central-limit-theorem-simplified-46ddefeb13f3>. Accessed January 2, 2021.
6. Editor MB. Choosing Between a Nonparametric Test and a Parametric Test. Minitab Blog. <https://blog.minitab.com/en/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test>. Published 2015.

Submitted 3/24/2021

III.2 The normal distribution and data analysis- Ben Seifer

Section III.1 of this guide describes the normal distribution and how it can affect decision making in data analysis. Statistical tests can be parametric or non-parametric where parametric tests assume the variable(s) being analyzed are normally distributed while non-parametric tests do not require this assumption. While this makes non-parametric tests more versatile, parametric tests have more power for detecting a statistically significant difference. Therefore, parametric tests are preferable if the analyzed variable(s) can be shown to have a normal distribution. In this section, we will describe typical methods for assessing the normality of a data set as well as touch upon how to approach non-normal data. All of the analysis demonstrated in this section can be performed using statistical software like Strata or SPSS, or with free programming languages like R or Python. Here, we used Python. For those interested in learning Python, the Dartmouth Library website has this excellent seven video resource (<https://researchguides.dartmouth.edu/pythonbites>).

Mean and Median

For a normally distributed data set, the mean should equal the median. If the percent difference between the mean and median is large, then data is likely skewed.

Example 1: You are collecting data from patients who presented to the hospital with an AKI and are interested in seeing if initial Cr correlates with length of hospital stay. The statistical test you want to use requires that Cr is normally distributed. The mean Cr of your data set is 2.0 and the median is 1.9. The percent difference is $1 - (1.9/2.0) * 100 = 5.0\%$. This is not a very high percent difference and the data set could very well be normally distributed, though further assessment is necessary.

Example 2: Your mean Cr is now 2.5 but your median is still 1.9. Percent difference is $1 - (1.9/2.5) = 24\%$. This is a relatively high percent difference and your data is likely skewed.

Standard Deviation

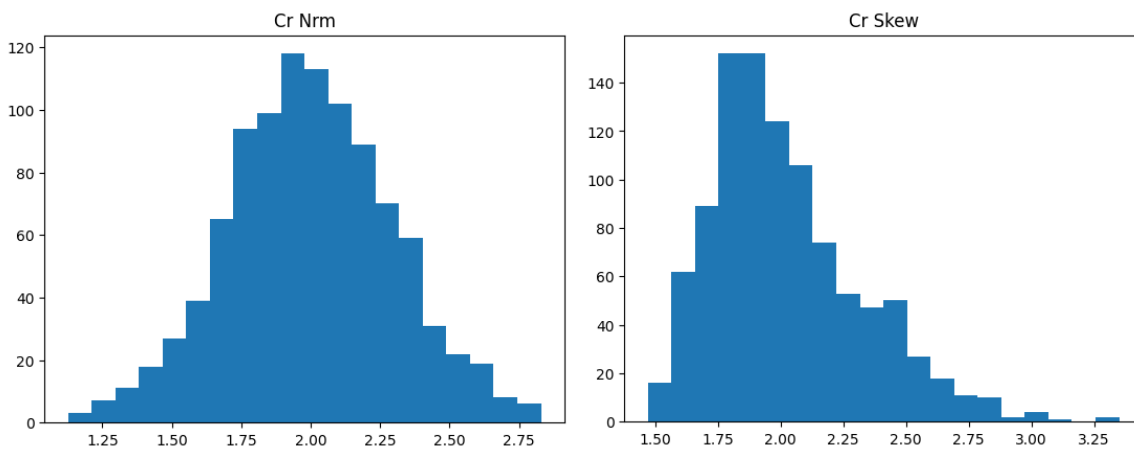
For a normally distributed data set, 95% of the data values should lie between -1.96 standard deviations and $+1.96$ standard deviations. Therefore, the 95% range of a data set can be estimated by multiplying the standard deviation by 2, then adding and subtracting that value to the mean to obtain the upper and lower value of the range respectively. If the estimated 95% range is within the range of the whole data set (i.e. between the minimum and maximum values) then this supports the assumption that the data set is normally distributed.

Example 1: In your data set with patients presenting to the hospital with AKI, the mean Cr is 2.0 and the standard deviation is 0.3. The maximum Cr is 2.8 and the minimum is 1.1. The minimum value of the estimated 95% is $2.0 - 2 * 0.3 = 1.4$, and the maximum value is $2.0 + 2 * 0.3 = 2.6$. The range of 1.4-2.6 is within the range of 1.1-2.8, supporting the assumption that the data set is normally distributed.

Example 2: Consider the same scenario but the minimum value of the data set is 1.5 and the maximum value is 3.3. The estimated 95% range of 1.4-2.6 is not within the range of the data set which is 1.5-3.3 indicating that the data set is likely skewed and not normally distributed.

Histograms, Q-Q Plots, and Box Plots

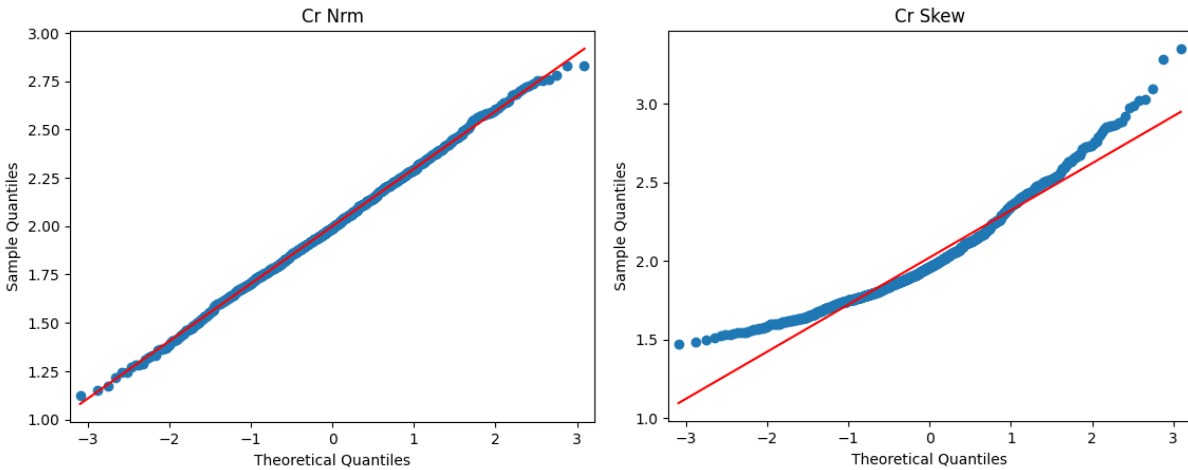
Visualizing the data set is an important step in assessing normality. A histogram of the data set allows the user to judge whether the distribution approximates a bell curve. Let's take a look at the histograms for the two data sets we just mentioned, one with Cr normally distributed and one with it skewed:



Note that these are the same data sets we were analyzing earlier and that they have equivalent means and standard deviations. The distribution on the left is convincing for a normal distribution. It appears fairly symmetric without a disproportionate tail on either side. The distribution on the right in contrast has a clearly larger tail on the right side indicating a right-sided skew.

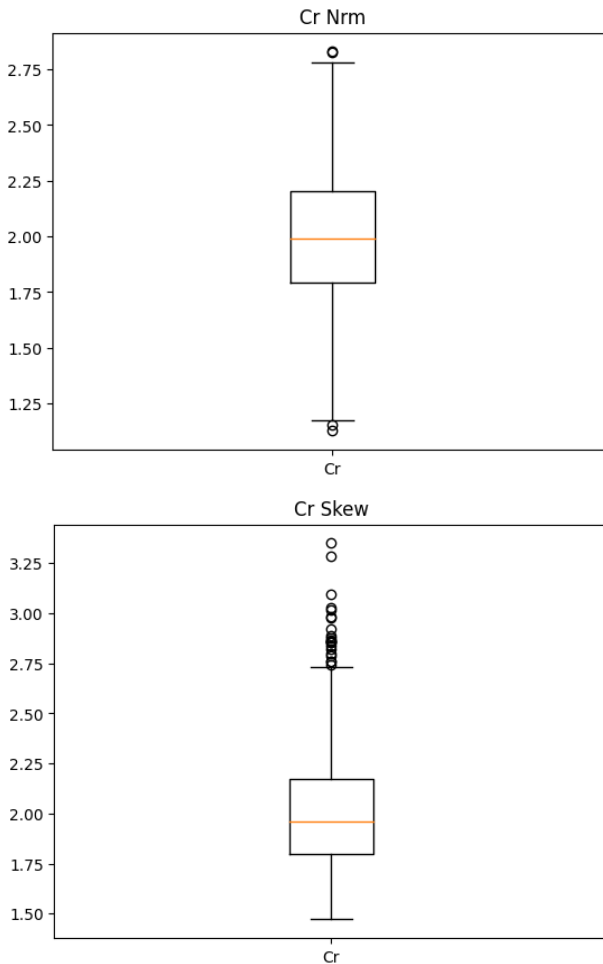
Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Another important, though less intuitive, data set visualization tool is the Q-Q plot. The Q-Q plot takes each data point and plots it against the value that would be expected if it came from a normal distribution with a mean of 0 and standard deviation of 1. The Q-Q plot should be linear along a 45° angle if the data is normally distributed. In the plots of our two Cr distributions below, we can see that the normally distributed data set follows this trend while the skewed data set does not:



The final visualization tool is the box plot. In addition to helping assess the symmetry of the distribution, the box plot is also useful for identifying extreme values in the data set. If a data set has too many extreme values, even if the distribution appears symmetric, this should raise concern that the data are not normally distributed. Take a look at the box plots for our two Cr distributions below:

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine



For all box plots, the horizontal line inside the box (the orange line in these two plots) represents the median. The vertical length of the box represents the inter-quartile range (IQR) with the lower line representing the 25th percentile and the upper line representing the 75th percentile. The two lines extending from the top and bottom of the box are called the “whiskers.” The whiskers are either the length of 1.5 times the IQR or extend to the minimum/maximum value, whichever is shorter. If there are values that exist past the ends of the whiskers, these are represented with circles and are considered to be extreme values (sometimes called “outliers”). In our box plot of the normally distributed Cr, we can see there are a few outliers, but for a data set containing 1000 data points, this is not very many and the distribution overall appears symmetrical. The box plot of the skewed distribution on the other hand is clearly not symmetrical, with the top whisker noticeably longer than the bottom one, and it has many extreme values.

Skewness

We have seen multiple ways of assessing for skewness in our data sets but is there a value that measures skewness directly? In fact, there are several different skewness coefficients that attempt to do this. Many statistical software packages will calculate the Fisher-Pearson coefficient of skewness and this can be obtained with Python as well. The equation for the Fisher-Pearson coefficient is quite

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

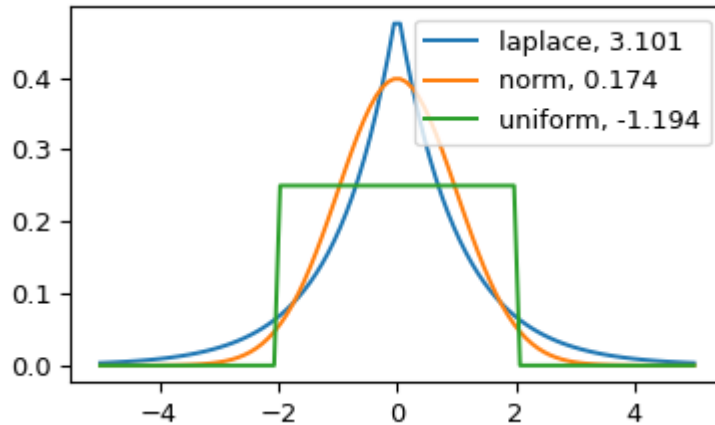
convoluted. What is important to know is that a normal distribution should have a coefficient of 0. A distribution that is right skewed will have a positive coefficient and a left skew will have a negative coefficient.

Our normally distributed Cr data set has a Fisher-Pearson coefficient of 0.02 and our skewed data set has a coefficient of 0.9. Therefore, both technically have a right skew. Of course, no data set will ever have a coefficient of exactly 0, and so the question becomes at what absolute value can we no longer consider the distribution to be normal? This depends on the size of the data set, with a Fisher-Pearson coefficient closer to 0 required for larger data sets. For a data set with $n = 1000$, as in our two data sets, the upper limit of the absolute value of the coefficient would be about 0.13. Our coefficient of 0.02 for the normally distributed Cr data set is well within this range. The coefficient of 0.9 is clearly outside of this range, but to give additional perspective, even if our data set was much smaller, say $n = 25$, a coefficient of $< \text{abs}(\pm 0.726)$ (i.e. < 0.726 for a right skew) would be recommended to assume the distribution is normal. So even for a much smaller data set, a coefficient of 0.9 would still be high, indicating how significantly skewed this data set is.

We can see that there is still some subjectivity to interpreting the Fisher-Pearson coefficient despite its attempt to quantify skewness. To this end, there is a statistical test available in most statistics software packages that generates a p value for a null hypothesis that the skewness of a given distribution is not different from that of a normal distribution. Applying this test to our two Cr distributions, the normally distributed Cr data has $p = 0.789$ indicating that it is very likely not skewed, and the skewed data has $p = 5.73 \times 10^{-25}$ indicating that it is almost certainly skewed.

Kurtosis

Kurtosis is another term like skewness that describes how a distribution differs from a normal distribution. Kurtosis describes how much of the data is concentrated around the mean. It is also referred to conversely as the “tailedness” of the distribution, or how much of the data is concentrated in the tails. A kurtosis of 0 means the distribution has no kurtosis compared to a normal distribution. Leptokurtic data has a greater concentration of data around the mean and less tailedness than a normal distribution and is represented by a positive kurtosis. Platykurtic data has less concentration around the mean and more tailedness and is represented by negative kurtosis. The figure below shows two types of distributions, a Laplace and uniform distribution, with a normal distribution to demonstrate kurtosis:



SciPv v1.10.1 Manual.

The Laplace distribution has more tailedness and less data around the mean than the normal distribution and has a positive kurtosis of 3.101. The uniform distribution has essentially no tailedness and highly concentrated values around the mean and has a negative kurtosis of -1.194. The normal distribution has a slight positive kurtosis of 0.174.

Again, interpreting these values, especially without visualization of the data set, does not give an obvious answer and can be somewhat subjective. As with skewness, there is a statistical test that can be used to generate a p value for kurtosis. This test generates a p value for a null hypothesis that a given distribution does not have more kurtosis than a normal distribution. Applying this test to the distributions above, we obtain $p = 2.27 \times 10^{-19}$ and $p = 1.44 \times 10^{-139}$ for the Laplace and uniform distributions respectively (both extremely significant for kurtosis) and $p = 0.556$ for the normal distribution (clearly not significant for kurtosis).

Critical Values of Skewness and Kurtosis

Skewness and Kurtosis can be divided by their respective standard errors to generate critical values that can also be used to judge how likely it is that a distribution is non-normal. You may be familiar with the equation for standard error (SE) of a mean:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Where σ is standard deviation and n is the sample size. The SEs of skewness and kurtosis similarly are functions of sample size but are more complex. They can be easier to demonstrate by using the equation for their squared value which equals the distribution variance (V):

$$SE_{skew}^2 = V_{skew} = \frac{6n(n-1)}{((n-2)(n+1)(n+3))}$$

$$SE_{kurtosis}^2 = V_{kurtosis} = \frac{4V_{skew}(n^2-1)}{((n-3)(n+5))}$$

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Note that it is best to calculate V_{skew} first as it can then be used in the $V_{kurtosis}$ equation. Let's go back to considering our Cr distributions. Both distributions have $n = 1000$ so SE_{skew} and $SE_{kurtosis}$ will be the same respectively for the two distributions:

$$SE_{skew}^2 = V_{skew} = \frac{6(1000)(1000 - 1)}{((1000 - 2)(1000 + 1)(1000 + 3))}$$

$$SE_{kurtosis}^2 = V_{kurtosis} = \frac{4V_{skew}(1000^2 - 1)}{(1000 - 3)(1000 + 5)}$$

$$SE_{skew}^2 = V_{skew} = \frac{6(1000)(1000 - 1)}{((1000 - 2)(1000 + 1)(1000 + 3))}$$

$$SE_{kurtosis}^2 = V_{kurtosis} = \frac{4V_{skew}(1000^2 - 1)}{(1000 - 3)(1000 + 5)}$$

$$SE_{skew}^2 = V_{skew} = 0.005982$$

$$SE_{kurtosis}^2 = V_{kurtosis} = \frac{4V_{skew}(1000^2 - 1)}{(1000 - 3)(1000 + 5)}$$

$$SE_{skew}^2 = V_{skew} = 0.005982$$

$$SE_{kurtosis}^2 = V_{kurtosis} = 0.02388$$

$$SE_{skew} = 0.07734$$

$$SE_{kurtosis} = 0.1545$$

We did not generate a kurtosis value for our Cr distributions but you will recall that skewness for the normal Cr distribution was 0.02 and for the skewed distribution was 0.9. Dividing these by SE_{skew} we get 0.3 and 11. These are our critical values of skewness. If these are within the range of -1.96 to 1.96, then it is likely that the distribution is not significantly skewed. We can see that 0.3 is very much within this range and 11 is quite outside this range.

Statistical Tests of Normality

There are two statistical tests available for assessing whether a given distribution differs significantly from a normal distribution: the Shapiro-Wilk (S-W) test and the Kolmogorov-Smirnov (K-S) test. The S-W test is the more powerful of the two and generates a p value for a null hypothesis that the distribution does not differ from a normal distribution. The K-S test is actually for comparing a given distribution to any desired distribution – the comparison distribution does not have to be a normal distribution, but it is frequently used in this way.

Running these two tests on our Cr distributions, the normal Cr distribution has $p = 0.658$ for the S-W test and $p = 0.573$ for the K-S test, and the skew Cr distribution has $p = 1.02 \times 10^{-18}$ for the S-W test and $p = 7.21 \times 10^{-4}$ for the K-S test. We can see that for both distributions, the S-W test produced the more certain result (more non-significant for Cr nrm and more significant for Cr skew). However, it is still useful to perform both tests as the determination of normality is most convincing if they both align.

Assessment of Normality Checklist

The checklist below summarizes the tools we have discussed and can be used to systematically analyze a data set for normality:

- Percent Difference between Mean and Median
- 2*STD Range compared to Data Set Range
- Histogram, Q-Q Plot, Box Plot
- Skewness, Kurtosis, and Critical Values
- S-W and K-S tests

The table below is taken from *Medical Statistics* by Barton and Peat, and shows how results from the different normality analyses can be juxtaposed to help decide if a data set can be assumed to be normal.

Table Summary of whether descriptive statistics and plots indicate a normal distribution

	Mean – median	Mean \pm 2 SD	Skewness and kurtosis	Critical values	K-S test	Plots	Overall decision
Birth weight	Probably	Yes	Yes	Yes	Yes	Probably	Yes
Gestational age	Yes	Yes	Yes	No	No	Probably	Yes
Length of stay	No	No	No	No	No	No	No

Peat and Barton. 2014.

We can see for the variable of Birth Weight that all analyses clearly or likely supported a normal distribution, and the overall decision (last column) was to consider this a normal distribution. For the variable of Length of Stay, all analyses clearly did not support a normal distribution and the overall decision was to consider this to not be a normal distribution. The variable of Gestational Age was not

as clear cut, having a few clear supporting analyses, a couple clearly not supportive analyses, and likely supportive plots. The overall decision here was to treat this as a normal distribution. The authors note that part of this decision was that the sample size for Gestation Age was > 100 which is a large enough sample size that parametric tests will be robust even if the distribution deviates somewhat from normal. They note that if the sample size were much smaller, “say less than 30”, then a parametric test would likely not be appropriate, at least prior to a transformation.

Data Transformations

Because parametric tests are more powerful than non-parametric tests, it may sometimes be worthwhile attempting to transform a non-normal data set to a normal distribution. Transforming data means applying a function to all data points such that a new value is generated for every data point.

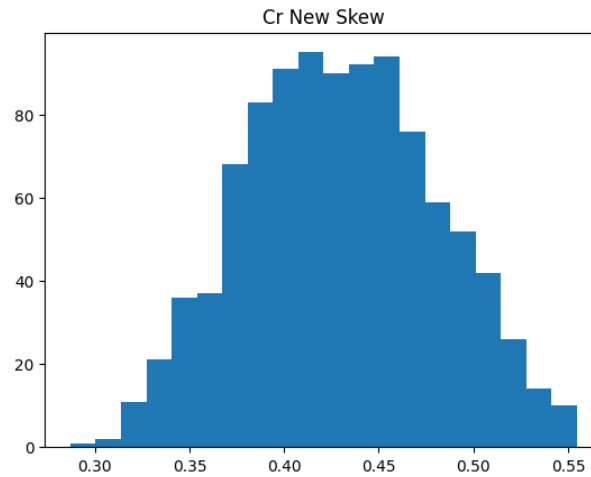
Some commonly used transformations include taking the square root of all data points, taking the base 10 or natural log of all data points, or taking the inverse (dividing 1 by the value) of all data points. However, the most systematic way of performing transformations is to use the Box-Cox transformation technique. The Box-Cox technique transforms the data multiple times with the goal of choosing the transformation that most closely approximates a normal distribution. This technique uses the equation below:

$$y'(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$

where y is the original data point value and y' is the new value. The recommended range for λ is usually -5 to 5. When λ is 0, $\ln(y)$ is used because y' would be infinite for all values in the top equation. Also note that this transformation covers a square root transformation (when $\lambda = 0.5$), an inverse transformation (when $\lambda = -1$), and a natural log transformation (when $\lambda = 0$).

Many statistical software packages include a Box-Cox function which both determines the most appropriate λ and transforms the data with that optimal λ . Applying the Box-Cox function in Python to our Cr skew data set, we obtain an optimal λ of -1.45 and a new histogram seen below:

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine



The skew seems to have disappeared! But when we apply the Shapiro-Wilk test, $p = 0.002$. This is much closer to non-significance than before the transformation but still is significant for a non-normal distribution. We find that for skew, $p = 0.627$, corroborating that transformed data is not skewed, but for kurtosis, we find that $p = 4.67 \times 10^{-6}$. This is why the S-W test is still significant – the transformed data is very kurtotic. Given this, we are probably best off using a non-parametric test on the original data set to compare the effect of Cr on length of stay. Note, that transformation of data will change the units of the variable and may make interpretation of results more challenging. This is another reason to potentially forgo an attempt at transformation and simply use a non-parametric test.

In summary, analysis of data variables for normal distribution is important to determine if parametric tests can be used for further statistical comparisons. Parametric tests are preferable to non-parametric tests as they allow for more power in detection of statistical significance. We have summarized a spectrum of tools that can be used to assess data for normality and have discussed how to use them together to decide if a parametric test can be used. We have also briefly discussed how non-normal data can be transformed to provide the option of using a parametric test. With this chapter in hand, the reader should be able to easily formulate a plan to analyze their data variables for normal distributions.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

References

1. Peat, J. K., & Barton, B. (2014). *Medical statistics: A guide to Spss, Data Analysis, and critical appraisal*. John Wiley & Sons Inc.
2. *Scipy.stats.boxcox#*. scipy.stats.boxcox - SciPy v1.10.1 Manual. (n.d.). Retrieved May 4, 2023, from <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.boxcox.html>
3. *Scipy.stats.kurtosis#*. scipy.stats.kurtosis - SciPy v1.10.1 Manual. (n.d.). Retrieved May 4, 2023, from <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kurtosis.html>
4. *Scipy.stats.skew#*. scipy.stats.skew - SciPy v1.10.1 Manual. (n.d.). Retrieved May 4, 2023, from <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html>
5. Standard errors of skewness and kurtosis are all the same for a set of variables. (2020, April 16). Retrieved May 4, 2023, from <https://www.ibm.com/support/pages/node/421769>

Submitted 5/5/2023

III.3 Understanding Odds Ratios and Relative Risk Ratios (Barry Howe)

(A) Fundamentals of OR/RR

OR and RR both tell you something about the risk of a bad thing happening, such as MI or catastrophic GI bleed. In general terms, an OR or RR that is greater than 1 indicates increased risk of the bad happening whereas less than 1 means a decreased risk.

For the difference between OR and RR, it's first helpful to remember the following (X being the disease or clinical event you are studying)

$$RR = \frac{\text{Probability of X happening if exposed}}{\text{Probability of X happening if not exposed and}}$$

$$OR = \frac{\text{Odds of X happening if exposed}}{\text{Odds of X happening if not exposed}}$$

However, Odds and Probability are very different...

$$\text{Probability} = \frac{\text{Number of times X happens}}{\text{Number of times X happens} + \text{number of times it doesn't}}$$

Probability is the event of interest divided by the total number of events. It's a percentage. If you roll dice, for instance, the probability of getting a 4 is 1/6. Odds, on the other hand is a different concept

$$\text{Odds} = \frac{\text{Number of times X happens}}{\text{Number of times X doesn't happen}}$$

The odds of rolling a dice and getting a 4 is 1/5

EBM application 1: odds ratio approximates RR when the event rate is very low

You can see pretty easily why folks will say that OR and RR are pretty similar when the event rate is low because odds and probability (which go into the formulas for OR and RR respectively) are effectively the same if you have a low event rate. If the event rate for X is 1 in every 1000, for instance, the probability of X is 1/1000 or 0.001 and the Odds of X is 1/999 or 0.001001...basically the same. On the other hand, OR and RR are quite different when the event rate is high.

(B) Study Design and the use of OR vs. RR

Imagine the rare condition of red nose in the smurf population. You wonder whether a daily glass of lemonade is associated with the risk of developing a red nose. You could pick out a population of smurfs without red nose and randomize them to receive daily lemonade or not (randomized trial) or you could just follow smurfs who drink daily lemonade and those who don't and see which ones develop red nose more frequently (prospective cohort) or you could go back and look at all smurfs in a sample, identify which were daily lemonade drinkers and which were not, then figure out which group had higher rates of red nose (retrospective cohort) or you could take a group of smurfs with red nose and compare it to a matched group of smurfs without red nose, then try to identify various factors—such as daily lemonade drinking--associated with the difference (case control).

Notice the easily confused difference between a case control study and a retrospective cohort: case controls separate groups by disease and then look back to identify exposures that may be associated with the disease; retrospective cohorts separate groups by exposure and then look back to identify whether disease is associated with those exposures. A cohort/RCT is thus able to follow the usual statistical procedures for obtaining a representative sample because there are no pre-determined limitations on how you get the sample (everyone either has the exposure or doesn't). However, a case control study by explicit design is avoiding the whole representative sample strategy and going straight for a concentrated sample of patients with the disease, then matching that sample to a group of controls.

EBM application 2: use OR in case control studies, RR in RCTs and Cohort trials

This distinction between case control and retrospective cohort discussed above is the key to understanding why you can't calculate relative risk for a case control study. Think of what you are asking with RR: what is the relative risk of developing red nose in a smurf who is a daily lemonade drinker? This question implies that you have taken a representative sample of lemonade drinkers, which you cannot claim to have obtained in a case control study.

In contrast to probability, however, odds do not imply you are making any statement of global likelihood (i.e., laying claim to a representative sample). You are just making a point estimate about a very specific set of sample data. Of course, if red noses in the smurf population is super rare, then the odds of exposure in a cases and controls likely approximates the probability of exposure.

III.4 Using Odds Ratio vs. a Hazard Ratio vs Relative Risk? (Erica Wadas)

Great question! You've just read when to use an Odds Ratio (OR) vs Relative Risk (RR) in this document (see previous). To cut to the chase – use OR in case control studies and RR in RCTs and Cohort Trials. You are not alone if you thought, "Wow I thought they were the same thing!" as when the disease is rare the OR and RR often have very similar numbers. Please see his section for the definition of OR and RR and when to use with great examples.

That leaves us with Hazard Ratios. Let's talk about when to use RR vs HR. Again, you might be saying, "Aren't these the same?" And again, you might be right. With some caveats...specifically the caveat of time.

A HR is used to compare a treatment group to a control group at a moment in time. When the HR is above 1 it means that the events of the treatment group are more likely to be seen. The event is whatever you chose, for example cure, adverse event, or death. Wait, this is starting to sound like RR isn't it?

But it's not. The most eloquent description of when and how to apply hazard ratios vs another was found here by Spruance et al called *Hazard Ratio in Clinical Trials* (Link: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC478551/>). They write:

"The Cox proportional hazards model is an appealing analytic method because it is both powerful and flexible. The hazard ratio, which is derived from this model, provides a statistical test of treatment efficacy and an estimate of relative risk of events of interest to clinicians. Examples of situations where the risk of an event is the question include the development of *Pneumocystis carinii* pneumonia in human immunodeficiency virus-infected patients, coronary reinfarction following stent placement, breast cancer in patients on estrogen supplements, and cardiovascular morbidity in patients taking aspirin.

However, the hazard ratio must be interpreted judiciously in clinical trials where the duration of events or the disease is the primary efficacy variable. The hazard ratio may be used for purposes of statistical hypothesis testing and as one indication of the amount of benefit (an increase in the odds of healing), but other measures must also be applied to understand the full importance of the study. Useful parameters on the time scale include the mean and median times as well as other percentiles to the study endpoint across treatment groups, and the median ratio."

To summarize and simplify, one should use hazard ratios when we wonder about the development of an event and then use a different model like relative risk when duration of event or disease duration is in question.

References:

1. Antimicrob Agents Chemother. 2004 Aug; 48(8): 2787–2792. doi: 10.1128/AAC.48.8.2787-2792.2004

Submitted 5/14/2018

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

III.5 Statistical Bias (Anthony Bambara)

Bias as it exists in evidence-based medicine is a term we use to describe systematic error that creates a deviation in the estimates produced by a study from the true parameters in the population. Bias in a study is due to a fault in the design of the study. Bias cannot be accounted for by simply increasing the size of the study sample but depending on the type of bias there are ways to mitigate the error created.

Selection bias: this represents the idea that the sample selected for a study will not be representative of the true population. Study results based on a sample not identical to the population are then not valid for use in the real world. There are many subtypes of selection bias that represent the different causes for a sample to deviate from the population.

1. **Self-selection/ volunteer bias:** This is the type of bias created when the sample is taken from people who are required to SEEK OUT the study. If a study were to be advertised without active recruiting the sample gathered would be comprised of those interested enough to volunteer. This group will likely have many differences from those who did not desire to volunteer, and some of those differences may make such a sample different from the general population in relevant ways. For example, if there was an offer of 100 dollars a week for people to volunteer for a study of effectiveness of new HTN drug, perhaps more unemployed people would volunteer, skewing the sample. Unemployed individuals might have a very different diet than the general population, a diet that could affect the blood pressure, thus making any conclusion about the new drug's effect on blood pressure not applicable to the general population.
2. **Non-response bias:** This is the idea that if some individuals choose not to respond to a survey or study offer and that their loss from the sample will skew the sample from the population it is intended to represent. For example, sending a letter to homes and asking people to drive to clinic to participate in the study will have a number of patients decide not to respond. That number may be higher among those not owning vehicles and so that portion of the population will be lost to the study.
3. **Under representation bias:** This term refers to the idea that if any group is left out of the sample it will create error in results achieved by a study from that sample because without that group the sample is not identical to the population. Both non-response and self-selection can cause underrepresentation bias as can recruitment for a study that simple does not reach certain groups in the population. Calling patients to ask for study participation will leave out those without a permanent phone number.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

4. **Survivorship bias:** This is the idea that if a sample is drawn from a group with a past event in common, then that group may have already undergone some self-selection and so could deviate from the general population. For example, if one study desired to measure cardiac function in population over 65 but took its sample from local nursing homes, it should be understood that those in nursing homes already likely failed independent living and so may be less cardiovascularly fit than the general population over 65, many of whom are more independent at home.

Solution to Selection bias: The goal of any study is to have a sample representative of the population. To best achieve this, it is important to carefully define your population, and then to randomly select members of that population to represent the whole. As people cannot be forced to participate there will always be some selection bias, but this will minimize the error created.

Measurement bias: This is the idea that systematic error in the results of a study can be created by the way data is collected or interpreted.

1. **Measurement Error:** This occurs when a device or technique used in data collection is skewed from the true value. If a study evaluates a new hematologic antigen test which will be confirmed by gold standard biopsy pathology in diagnosis of a cancer, then it is important that the measurements by pathologist are accurate for the results to be valid. If our pathologist rushes and there is a percentage of cancers missed on biopsy, then the percent found by the antigen assay might be artificially inflated and our conclusions will deviate from the truth.
2. **Observer Bias:** The phenomenon that occurs when the observer or researcher interprets or records events other than they are intentionally or unintentionally because of expectation for a certain outcome.
3. **Recall bias:** The idea that subjects of a study will tend to remember things differently than they truly happened. In a study where patients were randomized to receive a drug to prevent chest pain and then asked about angina episodes patients paying more attention might report more chest pain than previous because of error in recall. Patients may also recall exposure differently depending on their experiences. For example, asking a mother of a healthy child if she had any dangerous exposures during pregnancy will result in a much shorter list than in a mother of a child with birth defects who now focuses her memory on what could have gone wrong.

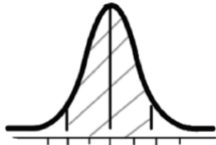
Minimizing measurement Bias: Though some types of measurement bias cannot be resolved observer bias at least can be taken out of the equation with the use of the blinding concept where observers/researchers do not know if a participant is assigned to treatment or control arm of a study and so intentional and unintentional differences in handling of data from these 2 groups is avoided.

Resources:

1. Bias in survey sampling. Stattrek.com. 2017.

September 2017

III.6 A Cartoon Introduction to Type I and Type II Error (Adam Eddington, GSM 4)



A Cartoon Introduction to Type I and Type II Error

-Adam Eddington GMS IV



HI everyone! The following is a cursory introduction designed to be a refresher on Type-I and Type-II error for visual inclined learners. But first, I want to welcome you all to San Lobos! A small-town famous for two things: All of its inhabitants are either office workers or lumberjacks and a local preponderance of werewolves.

Yes, unfortunately this little hamlet is beset by an endemic lycanthropy.

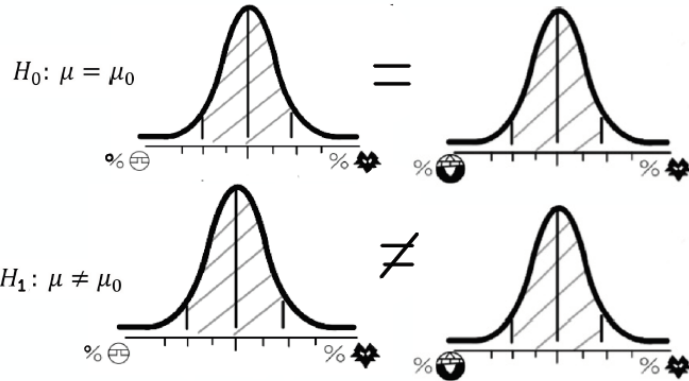


Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine

As nearly all individuals have some degree of the malady, we've defined disease progression along a continuous spectrum from 0-100% werewolf.

I've noted the work environments of the two groups are rather distinct. I hypothesize that compared to office workers; lumberjacks have a different mean degree of werewolfism.

In comparing the mean werewolfism between the office workers and the lumberjacks, we are trying to see if these populations differ from each other. The hypothesis that there is no difference between them is called the null hypothesis often abbreviated as H_0 .

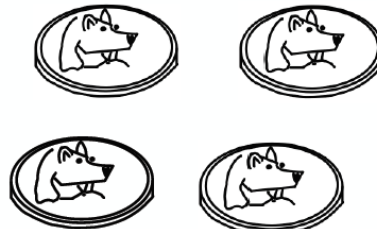


The hypothesis that there IS a difference is called the theoretical hypothesis or H_1 .



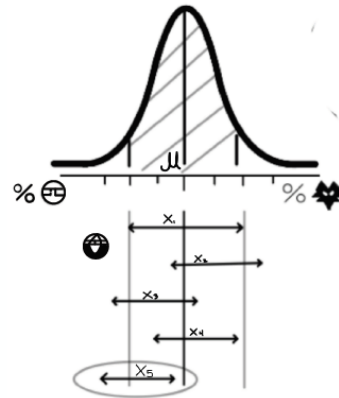
But what do we mean by "difference"? Logically, any two groups will have some differences between them. Consequently, we set limits and within them we regard the samples as not having any **significant** difference.

When we set the limits at twice the standard error of the difference and regard a mean outside this range as coming from another population, we will, on average, be wrong about 1 time in 20 if the null hypothesis is in fact true. This has nearly the same probability as tossing a coin 4 times and getting the same face each time (6.3%). If we obtain a mean difference bigger than two standard errors, there are only two explanations: either an unusual event has happened, or the null hypothesis is incorrect.



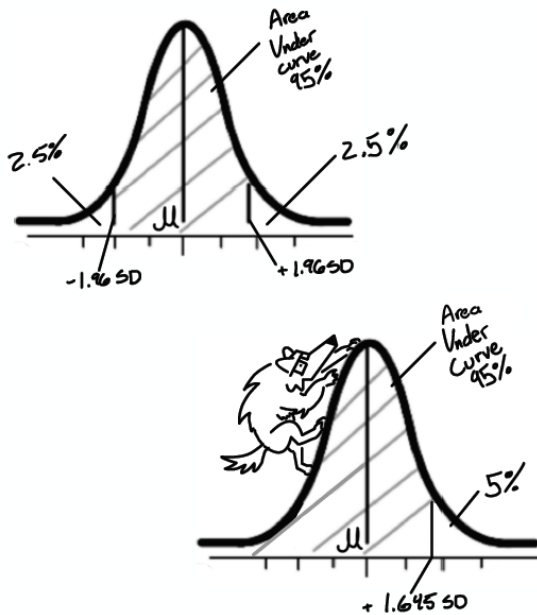
Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine

To reject the null hypothesis, when it is true, is to make a type I error or false positive. The level at which a result is declared significant is known as the type I error rate, this percentage is often denoted by α . In this example, it is set at 0.05. Another way to describe it is to say that we hypothesize that 95% of the sample means taken are projected to be in this shaded region.



If we take sample 1 and see that it lines up exactly with the hypothetical range, we would fail to reject the null hypothesis. Sample 2 is shifted but its standard of error still covers the hypothetical mean (μ). If we look at all examples in aggregate, we can see that, in this case, we SHOULD fail to reject the null hypothesis.

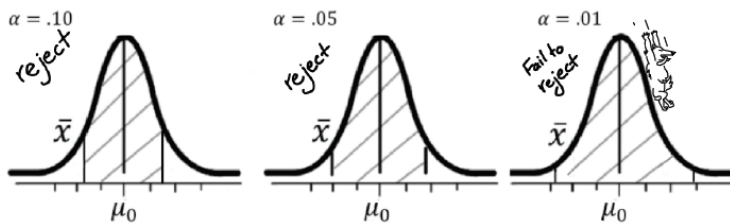
Now imagine that we only took sample 5. This would result in us incorrectly rejecting the null hypothesis and commit a type-1 error.



When researchers want to make a hypothesis, they can simply state that two groups have different means, or they can hypothesize the directionality of the mean. For example, when we say that we believe the mean percentage of werewolfism among lumberjacks is different than office workers, we aren't making a comment on HOW they are different (who has a higher or lower rate) simply that they ARE.

In a non-directional study, the cutoffs for statistical significance are placed at both ends or "tails" of a normal distribution. This means that if the two groups differ from one another by at least 1.96 SD, a nondirectional hypothesis would be supported.

A directional hypothesis moves that cutoff for statistical significance to + or - 1.645 SD (note that both cutoffs have the same total area under the curve, 5%) depending on the direction of the hypothesis. An example would be if we hypothesized werewolf prevalence among lumberjacks was HIGHER or LOWER than their cubicle-bound counterparts.

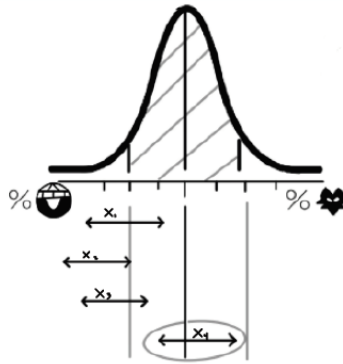


We can adjust α and therefore change the chance of Type I error. As α decreases the critical value to reject the null hypothesis moves outward and will be able to "catch" more sample means. This is analogous to widening a basketball hoop, the wider it is the more balls go in. But at what cost? This outward move of the critical values may also "catch" a mean from a DIFFERENT population. In that case we would fail to reject the null when it would be the correct thing to do. Therefore, we can see that the chance of a Type II error increases as α increases.

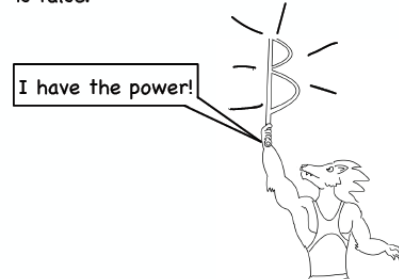
Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine

Type II error is failing to reject a false null hypothesis. Here we can see that the sample average mean shows a rejection of the null hypothesis. If we look at the samples in aggregate, it is clear we SHOULD reject the null.

However, if we only took sample 4, we would incorrectly accept a false null hypothesis and thus commit a type II error.



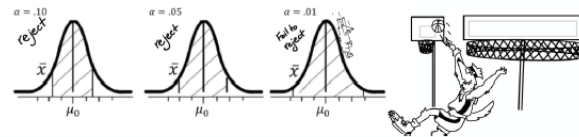
Type II error rate is often denoted as β . The power of a study is defined as $1 - \beta$ and is the probability of rejecting the null hypothesis when it is false.



The power of a study can be influenced by the following factors.

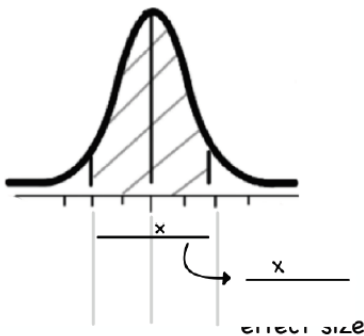
Significance (alpha changes)

Inversely related to power as explained above.



$$\alpha = \beta$$

The Effect Size



The relationship between effect size and type II error is fairly simple. The larger the effect size the less likely you will incorrectly sample a mean similar to your control. Put another way, the larger the effect being measured the more dramatically the sample mean will shift.

A useful illustration of the influence of effect size is a sample size formula for a two-sided, two group parallel trial with a continuous outcome with α of 5% and β of 20% (standard values are α of 5% and β of 10% or 20%).

$$n = 16\sigma^2/d^2$$

variability of the data

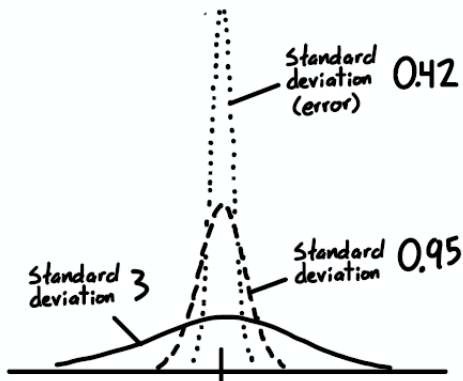
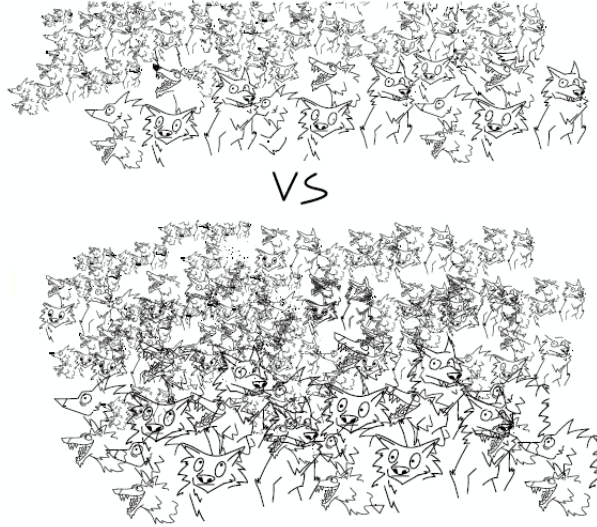
effect size



We see here that with all other factors being held constant, the sample size goes down inversely as the square of the effect size.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

For example, if we hypothesized that the average percent difference in werewolfism between groups was 4% and the between subject's standard deviation is 10% we would require $n = 16 \times 100/16 = 100$. However, if we suspected the difference between groups to be 5%, we would require $n = 16 \times 100/25 = 64$. So we see that the larger the difference the smaller the necessary sample size.



Sample Size

A larger sample helps to reduce standard deviation. This makes some intuitive sense as the larger the population, the more end up in the center of a bell curve therefore moving the standard deviations inward. This makes differences easier to see even if they are small.

In practice, the sample size is often fixed by restraints, such as finance or resources, and calculations are used to determine what an effect size would realistically have to be for statistical significance to be measurable. If this is too large, then the study will have to be abandoned or increased in size.

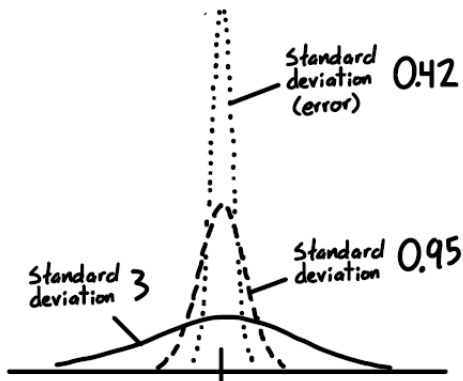
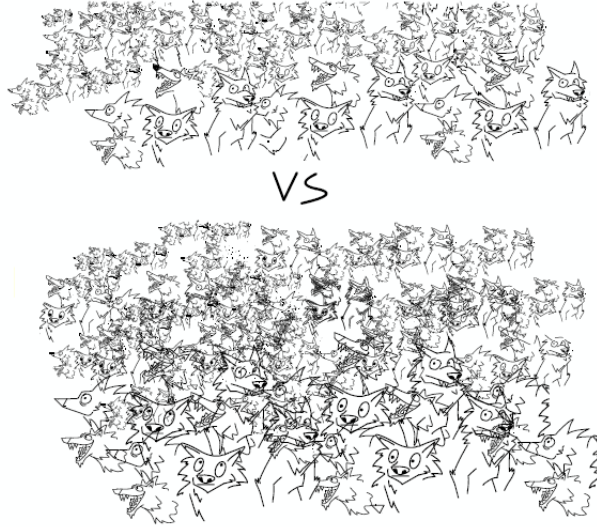


We have 12 sick patients, \$34.74, and this stinky. What can we show with that?

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

For example, if we hypothesized that the average percent difference in werewolfism between groups was 4% and the between subject's standard deviation is 10% we would require $n = 16 \times 100/16 = 100$. However, if we suspected the difference between groups to be 5%, we would require $n = 16 \times 100/25 = 64$. So we see that the larger the difference the smaller the necessary sample size.



Sample Size

A larger sample helps to reduce standard deviation. This makes some intuitive sense as the larger the population, the more end up in the center of a bell curve therefore moving the standard deviations inward. This makes differences easier to see even if they are small.

In practice, the sample size is often fixed by restraints, such as finance or resources, and calculations are used to determine what an effect size would realistically have to be for statistical significance to be measurable. If this is too large, then the study will have to be abandoned or increased in size.



We have 12 sick patients, \$34.74, and this slinky. What can we show with that?

Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine

In summation, statistical tests are just like diagnostic tests in that we can get false positives which is rejecting a null hypothesis when it is "true". This is a Type I error.



We can also get false negatives which is failing to reject the null even when it is not true. This is a type II error.

		Actual Condition	
		H_0 "True"	H_1 "True"
Study Conclusion	Do not reject H_0	Mean correctly inside nonrejection region	Type II error β
	Reject H_0	Type 1 error α	Mean correctly outside rejection region $1-\beta$

A solid understanding of these errors, the factors that influence them, and how they impact studies, can help authors and readers appreciate the impact of the data. For a more thorough (and less hairy) review of these concepts please see the following EBM handbook chapter 14.

Key

- H_0 - null hypothesis
- H_1 - theoretical hypothesis
- $1-\beta$ - Power
- α - type I error rate
- β -type I error rate

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

References:

1. Differences between means: Type I and Type II errors and power. (2020, October 28). Retrieved December 3, 2020 from <https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/5-differences-between-means-type-i-an>
2. Logic of Hypothesis testing Errors. (n.d.). Retrieved December 3, 2020 from http://onlinestatbook.com/2/logic_of_hypothesis_testing/errors.html
3. To Err is Human: What are Type I and II Errors? (2020, March 04). Retrieved December 4, 2020 from <https://www.statisticssolutions.com/to-err-is-human-what-are-type-i-and-ii-errors/>
4. AltmanDG. Comparability of Randomised Groups. Journal of the Royal Statistical Society. Series D (The Statistician). 1985;34(1);125-136. <https://www.jstor.org>

III.7 Statistical Error (Brenton Nash, Lizzy Schink, comments by Ken Phelps)

Hypotheses

Whenever a statistical test is being performed to compare two groups, investigators are testing hypotheses. These hypotheses may not always be explicitly stated but they exist in principle all the same. They are the research/maintained/alternative (H_1) and the null (H_0) hypotheses.

Let us provide an example of the research and null hypotheses. Let us say we expect more emergency room visits on the night of a full moon. We would state our hypotheses thusly...

H_1 : There are more emergency room visits on nights during which there is a full moon compared to nights on which there is not.

H_0 : There is no increase in emergency room visits on nights with a full moon compared to nights on which there is not.

The research hypothesis is basically the research question the investigators are asking when they perform as statistical test to compare to groups. The null hypothesis is commonly referred to as being the opposite of the research hypothesis. BE CAREFUL with this definition – “opposite” is a bit of a misnomer.

If the null hypothesis were the exact opposite of the research hypothesis, instead of saying that emergency room visits do not increase with full moons, one might say that emergency room visits decrease with full moons. This is not quite the same. This is saying that the research hypothesis would not be supported if there was a statistically significant decrease in emergency room visits on nights with a full moon. In actuality, the investigators would want to know if there was no statistically significant movement either way (no change).

It is in light of this potential confusion that this author prefers to define the null hypothesis as a statement that negates the research hypothesis.

Directionality/“Tailed-ness”

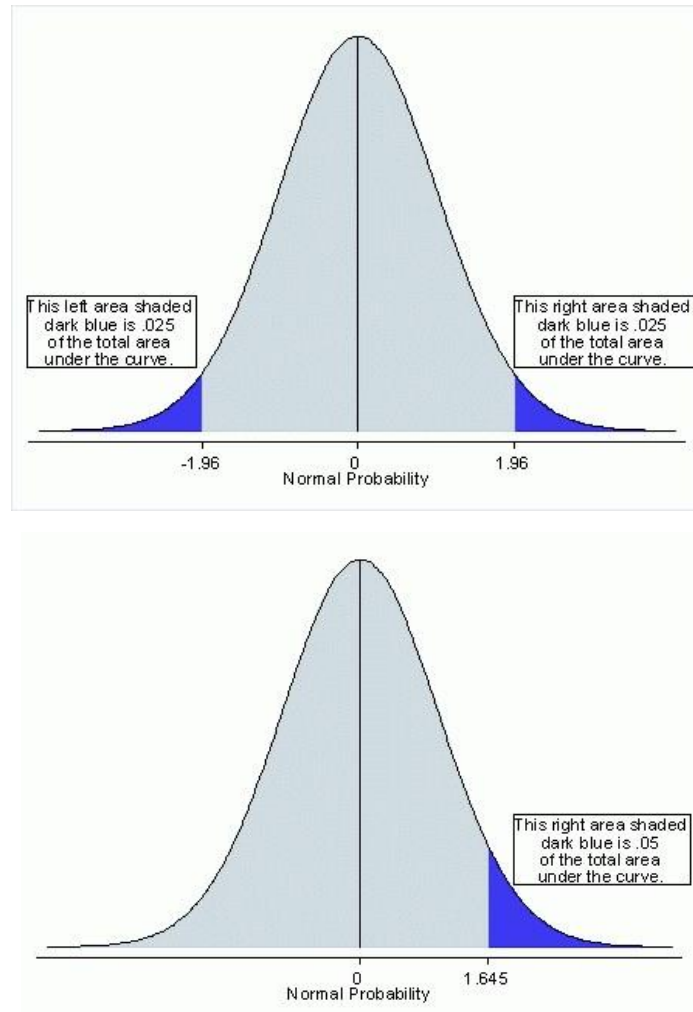
When researchers make hypotheses, they can either hypothesize on the nature of the relationship between the two groups tested (one group mean being higher or lower than the other) or they can simply hypothesize that the two groups are different.

Directional Hypothesis: There are more emergency room visits on nights during which there is a full moon compared to nights on which there is not.

Non-Directional Hypothesis: The number of emergency room visits differ on nights during which there is a full moon compared to nights on which there is not.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Ideally, the directionality of the hypothesis would determine the nature of the statistical analysis. (NOTE: in medical research it is not uncommon for a directional hypothesis to be tested with a non-directional statistical test). A directional hypothesis would ideally be tested with a one-tailed test. A non-directional hypothesis would ideally be tested with a two-tailed test. The “tailed-ness” of the tests refers to the placement of the cut off for statistical significance. Refer to the figure below.



These figures were obtained from UCLA’s IDRE website. The first figure shows that the cutoffs for statistical significance for a non-directional hypothesis are placed at both ends or “tails” of a normal distribution meaning that if the two groups differ from one another by at least ± 1.96 SD, a non-directional hypothesis would be supported. The second figure shows that the cutoff for statistical significance for a directional hypothesis is placed at either end or “tail” of a normal distribution meaning that if the two groups differ from one another by either + or $- 1.645$ (depending upon the direction of the hypothesis) a directional hypothesis would be supported.

Hypothesis Testing Parlance

In case you ever have a discussion with a professional statistician regarding hypothesis testing, it is helpful to remember the following bit of semantics.

Research hypotheses are either supported or not supported by the data, they cannot be proven or true. This is because our statistical tests only ever provide evidence for our hypothesis, they cannot prove that something exists.

Null hypotheses are either rejected or failed to be rejected. They are not generally referred to as supported and they are never proven nor true.

Statistical Error

Statistical tests can result in false positives and false negatives (like diagnostic tests).

	Null Hypothesis is True	Research Hypothesis is True
Null Hypothesis is Rejected	Type I Error $p = \alpha$	Correct decision $p = 1 - \alpha$
Failure to Reject Null Hypothesis	Correct decision $p = 1 - \beta$	Type II Error $p = \beta$

NOTE: that “true” in this section refers to a theoretical, unknown truth. One should never refer to their research or null hypotheses as true or false.

From the table hopefully, you can see that a false positive is a Type I Error and a false negative is a Type II Error.

Type I Error: false positive. Incorrectly rejecting a true null hypothesis. Consequently, there would be a claim of support for a false research hypothesis.

Type II Error: false negative. Incorrectly failing to reject a false null hypothesis. Consequently, there would be a claim of no support for a true research hypothesis.

(Enough double negatives for you?)

The α value with which we are all familiar is your probability of making a type I error (false positive error). By convention, α is commonly set at 0.05. This means that there is a 5% likelihood, assuming the data you are analyzing occurs on a normal distribution, that your result was obtained merely by chance. If your results were due solely to chance, then there would likely be no *true* difference between the two groups.

It should be easy to see that $1 - \alpha$ is our ability to avoid a type I error. Thus, with the same 0.05 set point, there is a 95% chance we are not making a type I error.

The β value is your probability of making a type II error (false negative error). If β were set at 0.10, there would be a 10% likelihood of making a type II error.

Power, or $1 - \beta$, is our ability to avoid type II errors.

What affects Type I and Type II errors?

Type I Error: the major determinant of a type I error is the threshold for statistical significance. Thus, if the p value of the study were set at 0.05 that the likelihood of a type I error would be 500 X that of a study in which the p value was set at 0.0001.

Additionally, type I error on a study wise basis is affected by the number of statistical tests performed upon the data. If investigators perform multiple statistical tests each with a p value of 0.05, the study wise level of significance decreases and is roughly equivalent to the number of tests times the level of significance for each test. This makes intuitive sense. If I perform two statistical tests on the same data set each with a p value of 0.05, the probability that I am obtaining my results solely due to chance is 10%. This concept is referred to as multiple tests or multiplicity.

Type II Error: the ability to avoid a type II error, or power, has three major determinants – effect size, level of significance, and sample size.

Effect Size: this should make intuitive sense. The larger the effect size, the easier it is to see the difference between two groups.

Level of Significance: if I set my p value at .01 instead of 0.05, it should be obvious that while I am less likely to make a type I error, I am more likely to make a type II error

Let us return to the example of emergency room visits and the lunar cycle. Let us say we perform the study, and we find that on average the emergency department has 100 visits per night on non-full moon nights and during a full moon the number increases to 105 visits per night. If our data set is highly variable, our standard deviation will be larger and, thus, it is possible that even though a difference between the two groups exists (105 is greater than 100) the averages do not differ by 1.645 standard deviations (a one-tailed test for significance at a level of $p < 0.05$). In fact, the two averages differ by only 1.644 standard deviations. If I set my threshold for significance to $p < 0.06$, I might have a positive study.

Recall the section on directionality and review the figures again. Notice that if research wants to maintain a study wise p value of < 0.05 for a two-tailed test, they must “split” the significance level between the two tails. That is the threshold for significance is now $\pm 2.5\%$. Consequently, two-tailed tests decrease power.

Sample Size: with a larger number data set it is often easier to see differences even if they are small. Think of this as turning a dial on a radio set. Increasing the sample size helps an investigator tune to the right channel and minimize the background static. This is because a larger sample helps to reduce variance. Recall that formula for standard deviation, in which N is in the denominator.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}}$$

(Contributed by Brenton Nash, March, 2018)

Another way to think about this is in trying to represent a clinical study design in a 2x2 table. The analogy to the attributes of a diagnostic test may prove useful.

(1 – beta) is the likelihood of finding the designated clinically significant difference between experimental subjects and controls if the author’s hypothesis is correct. Alpha is the likelihood of finding that difference if the null hypothesis is incorrect. If we view the trial as a test, 1 – beta is the sensitivity, and 1 – alpha is the specificity of the trial. Sensitivity in the context of test performance is analogous to *power* in trial design.

Trial result	Author’s hypothesis (H₁) is correct	Null hypothesis (H₀) is correct
Positive (shows a significant difference between experimental group and controls)	Probability (1 – beta) of a true positive trial result (sensitivity or power)	Probability (alpha) of false positive trial result
Negative (no significant difference between experimental group and controls)	Probability (beta) of false negative trial result.	Probability (1 – alpha) of a true negative trial result (specificity)

Studies are *designed* to incur risk beta that a designated, clinically significant difference will not be found even though it is present, and risk alpha that the difference will be found even though it is not present. Conventionally, the designated values for alpha, beta, and power are 0.05, 0.2, and 0.8, respectively. The values, in combination with hypothesized means and standard deviations in experimental subjects and controls, determine the number of patients that must be studied.

In the 1980s, Diamond and Forrester suggested that alpha and beta are not the whole story. As with diagnostic testing, they argued that Bayesian analysis should be applied to trial results exactly as it is applied to test results. To interpret a positive or negative test result quantitatively, we need to know the prior probability of the disease in question (i.e., probability of the disease before the test was performed). The analogous principle is that authors and readers must estimate *the prior probability that the hypothesis was correct* in order to interpret the result of a clinical trial. This logical step is rarely if ever taken.

For example, assume sensitivity of 0.8 and specificity of 0.95 for a test to detect disease x. Assume also that the prior probability of x before the test is performed is 0.1. Bayes’ theorem yields the following 2x2 table. The numbers in the cells are the probabilities of each outcome given our estimates of sensitivity, specificity, and prior probability of disease.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Test result	Disease present (p = 0.1)	Disease absent (p = 0.9)
Positive	0.08	0.05 (0.045 rounded up)
Negative	0.02	.85

In this example, the likelihood of disease x given a positive test (positive predictive value) is 0.08/0.13, or approximately 0.6.

The analogy to interpretation of a clinical trial is straightforward:

Trial result	Author's hypothesis (H ₁) is correct (p = 0.1)	Null hypothesis (H ₀) is correct (p = 0.9)
Positive	0.08	0.05 (0.045 rounded up)
Negative	0.02	.85

The positive predictive value of a positive trial result is 0.6. In other words, given a positive result, the likelihood that the author's hypothesis is correct is 0.6, and the likelihood that it is incorrect is 0.4.

One lesson: Only investigate hypotheses that are very likely to be true.

Contributed by Ken Phelps, MD, Albany Medical College, January 2020

Decoding the “adjusted” analysis: an exercise in the avoidance of type II error

Just as confounding variables can demonstrate spurious relationships, suggesting they exist when they do not (type I error), they can also obfuscate true relationships (type II error). Although we've been trained to fear type I error above all else (with good reason), one might argue that, especially in the setting of clinical medicine with potential lives at stake, systematically committing type II error can be just as egregious. In many instances, failure to deem a study as “practice-changing” can cause as much harm as changing a practice in accordance with faulty data. An extreme example of this is seen in experimental trials of cancer drugs, where clinicians might choose to use a promising therapy. Such a clinician may tolerate a p of 0.10 when n = few and the outcome is a response in treatment-refractory AML. Keeping our p's < 0.05, the rest of us can reduce our chance of committing type II error by considering adjusted data. In order to see how, let's first go back to the basics.

When one calculates almost any test of statistical significance (e.g., Z score, chi square), the formula is some iteration of:

$$\text{Statistic} = \text{difference between groups} / \text{error}$$

Although the nitty gritty of what goes in the numerator and denominator varies, the idea is consistent: the magnitude of the difference between groups with respect to the outcome of interest is in the numerator, error in the denominator. If our study is designed correctly, the numerator should reflect the amount of change produced by the independent variable (e.g., intervention in an RCT). The denominator should reflect the amount of difference between groups that might be expected due “trivial” variation among individuals (we are all unique after all). In RCT’s, this denominator value is the difference due to chance. In any such statistic, the goal of the calculation is to put the difference between groups (numerator) into context of error (denominator) in order to determine whether the groups are “truly” different.

With this framework in mind, let’s turn to the data that we actually enter into our statistic equation. The equation can be run using either raw data or what is called “adjusted” data. In essence, when raw data are used, we make the assumption that *everything* in the denominator should really be there (i.e., everything is trivial). On the surface this makes sense, especially in the setting of random assignment, where any difference between the groups is due to chance alone.

However, by only using raw data, we ignore the fact that the denominator doesn’t only capture differences due to “trivial” variation. The denominator also captures between-group differences with respect to other variables may impact the outcome of interest. Occasionally, these other variables have even been studied and their impact is known. If the groups are “different enough” with respect to any of these other variables (even if this difference is not technically significant, more on that below), this may inadvertently bias the study outcome.

Sounds important. How do we adjust then? “Adjustment” basically consists of 4 steps:

1. Choose one or more of these “other variables” (officially called “covariates”) and measure them in all patients at baseline
2. Calculate the difference between the groups with respect to the covariates
3. Calculate the expected effect of the covariate on the outcome of interest
4. Remove the contribution of this expected effect from the statistic

The 4th step essentially removes (“adjusts” for) the contribution of the covariate(s) to the error term (denominator). In effect, adjustment takes some of what we mislabeled as, “trivial” variation when we were using raw data, and re-labels it as, “expected variation based on between-groups differences in one or more important covariates.” In this way, you can think of adjustment as decreasing the magnitude of the denominator in our statistic equation. In other words, adjustment is able to eliminate some of the “noise” due to differences in important covariates (note that this is not actually noise) in the same way increasing sample size eliminates “noise” due to chance differences among individuals. Without changing the cutoff value (i.e., p is still < 0.05), adjustment increases the power of the study to detect a difference when one truly exists (in a way that is no less “legit” than increasing the sample size of a well-designed study!). One could see why adjustment is especially useful when power is limited by sample size.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

It's easy to see why non-randomized studies, such as cohort studies, are adjusted for potential confounders (analogous to the "covariates" described above). But at this point, you may be thinking, "Hey! The whole point of randomization is that these baseline differences are randomly balanced, so how could any of this matter? Who do these authors of randomized studies think they are pulling a stunt like adjusting for covariates?"

Much to your and my chagrin, even properly randomized groups are almost always different. Although proper randomization ensures that any difference between the groups is due to chance alone, it does not guarantee against baseline differences factors that may have strong prognostic implications, even in relatively large trials. Further, these differences in baseline factors, even when non-significant, can strongly influence the outcome of a trial (Altman, 1985). This is an important point. For example, especially in studies of survival, even non-significant differences in age between groups (e.g., $p = 0.10$) could strongly influence an outcome.

When the authors of a study conduct a significance test to compare the groups and show no significant difference in baseline variables ($p < 0.05$), this is *not* grounds to move forward without further consideration of these variables. In fact, the idea behind the p value is to determine the probability that an observed difference occurred due to chance. In a randomized trial, any observed difference is *necessarily* due to chance. As Altman (1985) eloquently summarizes, "performing a significance test to compare baseline variables is to assess the probability of something having occurred by chance when we know that it did occur by chance. Such a procedure is clearly absurd." He goes on to explain that, "it is the strength of the association rather than the significance level (which also depends upon sample size) which is of importance." Adjustment is the statistical method of incorporating this "strength of association" when accounting for baseline differences in covariates. Clearly, then, adjustment should be favored over "eyeballing" a list of p-values comparing the baseline characteristics of the groups. For further reading on this subject including specific examples, please see the paper by Altman.

Hopefully, this discussion debunks the (false) idea that, if the association between independent and dependent variables is strong enough, the raw data will be significant. If there are good reasons for adjusting for covariates, * the adjusted data will actually remove the tendency of otherwise low powered trials to commit type II error. Most importantly, it will do this without inflation of type I error (Hernández et al., 2004; Kahan et al., 2014). This is possible because, if the adjustment for prognostic factors affects the overall comparison, it is equally likely to do so in either direction (Altman, 1985). This idea has been supported when applied to a wide variety of study types, including RCT's with dichotomous and continuous outcomes (Hernández et al., 2004; Canner, 1991; Raab et al., 2000).

*Good reasons for adjusting for covariates: (1) a relationship between the covariate and the outcome has been demonstrated in previous studies, or (2) the author makes a compelling case for adjusting for no more than one or two unstudied covariates (see Lee, 2016). Although one should be skeptical when an author simply adjusts without providing an explanation, adjustment for a reasonable covariate is generally more beneficial than harmful. For more on which covariates to adjust for, see Raab et al. (2000).

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

References:

1. Lee PH. Covariate adjustments in randomized controlled trials increased study power and reduced biasedness of effect size estimation. *J Clin Epidemiol.* 2016;76:137-46.
2. Hernández AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol.* 2004;57(5):454-60.
3. Kahan BC, Jairath V, Doré CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials.* 2014;15:139.
4. Canner PL. Covariate adjustment of treatment effects in clinical trials. *Control Clin Trials.* 1991;12(3):359-66.
5. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. *Control Clin Trials.* 2000;21(4):330-42.
6. Altman DG. Comparability of Randomised Groups. *Journal of the Royal Statistical Society. Series D (The Statistician).* 1985;34(1):125-136. <https://www.jstor.org/stable/pdf/2987510.pdf>
7. Contributed by Lizzy Schink, Dartmouth College (Geisel School Yr IV), April 2018

III.8 Statistical Significance- not as simple as $p < 0.05$ (Luke Mayer, GSM4)

There is a rising chorus of voices advocating and even petitioning for the removal of the term "statistical significance" from use in the scientific literature. There are rational arguments for and against this proposal; this chapter will be an effort to explore both sides.

For the removal of "statistical significance" from the literature.

Amrhein et al. (1) describe a pervasive misunderstanding of statistical process that has "warped" the scientific literature. They describe an epidemic of researchers who fail to thoughtfully present their results. To highlight their point, they raised concerns about researchers concluding that there is "no difference, or no association just because a P value is larger than the threshold such as 0.05". They pose, as an example, a study that demonstrates a risk ratio of 1.2, with a 95% confidence interval of .97 to 1.48 ($P=0.091$). The second hypothetical study, whose goal is to verify the results of the first, also demonstrates a risk ratio of 1.2. This study, however, was more precise, with a 95% confidence interval of 1.09 to 1.33 ($P=0.0003$). Amrhein et al. note that "it is ludicrous to conclude that the statistically nonsignificant results [of the first study] showed "no association", when the interval estimate included serious risk increases; it is equally absurd to claim these results were in contrast with the [second study's] results showing an identical observed effect." A Nature study of 791 articles across 5 journals found that 51% misrepresented their statistical findings, suggesting that 'non-significance means no effect'. Amrhein et al. proposed a ban on the term "statistical significance", and on the very idea of an arbitrary cutoff being used to define significance (for example: $P = 0.05$). The group received 800 signatures supporting their proposal within a week of its initial presentation. The group does, however, endorse the ongoing use of p-values, confidence intervals, and other statistical measures. They simply denounce the dichotomization of values as either 'significant', or 'not significant'. They argue that nature is full of nuance, and that the effort to dichotomize results into either significance or irrelevance is not only reductive, but also futile.

Against the removal of "statistical significance" from the literature.

Ioannidis et al. disagree. Where Amrhein et al. denounced the reduction of a complex set of findings to a dichotomous outcome, Ioannidis et al. point out that "dichotomous decisions are the rule in medicine and public health interventions. Any intervention, such as a new drug, will either be licensed or not and will either be used or not" (2). They argue that "significance (not just statistical) is essential both for science and for science-based action, and some filtering process is useful to avoid drowning in noise." Ioannidis's group acknowledges that the proverbial carrot of statistical significance will lure careless or less-than-scrupulous researchers into cherry-picking data or utilizing poor methodology in order to get significant results. They write, "absent pre-specified rules, most research designs and analyses have enough leeway to manipulate the data and hack the results to claim important signals." They even referenced a survey "completed by 390 consulting statisticians" that found "a large percent perceived that they had received inappropriate requests from investigators to analyze data in ways that obtain desirable results". Despite this problem, they suggest that the wholesale ban of the term 'statistical significance' would be "overturning the tables", and that it would be more prudent to try "to fix what is

lacking and set better and clearer rules". They argue that statistical analysis should be expected to be fully, rigorously pre-thought and documented, so as to avoid the pitfall of tailoring analytic methodology to produce a certain result. They argue, "more thought should go into research before it is conducted, not after the data has been inspected." They propose that researchers be expected to make their raw data public in an effort to enhance trust. Finally, they suggest that statistical illiteracy is the root cause of the problems raised by Amrhein et al. They suggest the ultimate solution, rather than a heavy-handed ban on 'statistical significance', ought to be an improvement in the statistical numeracy of the scientific workforce.

Actionable approaches for the reader?

One might wonder how they might become a more responsible consumer of the literature in light of the arguments laid out above. This author humbly suggests three things the consumer (patient, physician, or other health professionals) should always consider.

First, if the result in question was presented as not statistically significant, one must remain aware that this does not necessarily mean that there is no relationship. The consumer ought to consider whether the study was adequately powered to parse the relationship in question. If the study was not adequately powered, a non-significant result could easily belie a true relationship.

Second, search for multiple sources. Rather than make changes to one's clinical practice based upon a single study, comb the literature for multiple sources. For example, multiple adequately powered studies failing to find a relationship, is certainly a strong foundation of evidence from which to make clinical decisions. However, if the literature is divided about the significance of a relationship, more nuanced thought is required to tailor one's clinical practice.

Third, remember to check if the research group adjusted their P values for how many statistical tests they ran. Without diving too deep, consider a study group that queries a database for 100 separate relationships. The group finds them all to be statistically significant relationships using a cutoff P value of 0.05. One could expect that ~5 of those significant findings were false positives. Considering that the likelihood of pulling up false-positive findings increases with the number of relationships a study queries, the group ought to address this and adjust their cutoff P values accordingly.

Finally, given the plethora of publications, the efficient and thoughtful clinician or patient may decide that using sources other than RCTs may solve the problems addressed above. The use of resources such as UpToDate, DynaMed, and ACCESS, as well as the many meta-analyses and systematic reviews to say nothing of Guidelines are tempting as sources that may "have done the work" for you. As always, there are risks and benefits to this strategy. Practicing EBM is a clearly a lifelong commitment.

References:

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019; 567(7748):305-307.
2. Ioannidis JPA. The importance of predefined rules and prespecified statistical analyses: do not abandon significance. *JAMA*. 2019;321(21):2067-8.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Submitted 4/6/20

III.9 Estimating Sample Size (Haley Moulton, GSM4)

Estimating sample size prior to a study is essential to report results with a given level of confidence. However, articles often don't report all the required parameters, and some don't report any sample size calculation at all.¹ Even if articles do report sample size, there can be a large difference between sample size used, and what the calculation would indicate should be the sample size. Yikes.

So how do you calculate sample size? And what does it really mean?

The Calculation:

There are a few different versions depending on the type of study, but this is for clinical trials where a study is looking at the effect of an intervention, comparing two groups with quantitative data²

$$\text{Sample size} = \frac{2SD^2(Z_{\alpha/2} + Z_{\beta})^2}{d^2}$$

Sample size = sample size **per treatment group**

SD = standard deviation

Z $\alpha/2$ = z-boundary of confidence interval, 1.96 for 95% CI

Z β = from Z table, 0.842 for 80% power

d = effect size = difference between mean values

Essentially this calculation is to estimate the minimum sample size necessary to produce 95% confidence that the sample mean of whatever outcome you're looking for is statistically and clinically significant. Sounds like an important thing to do before starting a trial, right?

Let's look at an example to give these variables some context. A randomized controlled trial by Lorenz et al.³ wanted to answer the question if using a suction mask during positive pressure ventilation in the delivery room for infants who do not establish effective spontaneous breathing reduces mask leak compared to conventional silicone masks.

A previous study on preterm infants in the delivery room had determined a mean facemask leak was 30%, and the standard deviation was 17%. They also performed a pilot study of the suction mask using a manikin model that reported a 95% decrease in leak using the suction mask. In testing this suction mask intervention with human subjects, they decided that a more conservative 50% decrease in leak would be clinically significant.

Now let us look at our variables to calculate sample size. **SD = 17** based on the previous study. **Z $\alpha/2$ = 1.96** for a 95% confidence interval. **Z β = 0.842** for 80% power. If 30% is the mean for conventional masks based on the previous study, a 50% decrease would mean 15% facemask leak for the silicone masks. The difference between means is then 30%-15%, so **d = 15**. Putting it all together we get:

Sample size = $\frac{2 \cdot 17^2 \cdot (1.96 + 0.842)^2}{15^2} = 20.14$. So, a sample size of at least 21 infants per group is needed.

Ultimately 45 infants were enrolled, 23 were randomized to the conventional mask, and 22 were randomized to the suction mask.

All else remaining constant, when the level of confidence is increased, you will need a larger sample size. The larger sample size ensures more sample means are within the given margin of error due to the fact that a larger sample size is more representative of the overall population.

Increase sample size → reduce standard error → sample more representative of population

Solving for the sample size requires knowing the population standard deviation, but that's something we generally don't know off the top of our heads so there are a few options to form an estimate:

- 1) Estimate from a previous study using the same population of interest (i.e., research replicating another study)
- 2) Conduct a pilot study to select a preliminary sample, and then use the sample standard deviation from the pilot study
- 3) If you have NOTHING else to go on, you can guess using the data range divided by 4. Not ideal.

Yes, the sample size calculation uses a few assumptions, but it must ALWAYS be calculated prior to a study otherwise you're assuming a whole lot more.⁴

Sample size is more than just reporting the number of participants. In researching evidence-based medicine, look for articles that not only mention sample size, but describe how they calculated it.

References/Footnotes:

- 1) Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. *BMJ*. 2009 May 12;338:b1732.
- 2) Charan J, Biswas T. How to calculate sample size for different study designs in medical research?. *Indian J Psychol Med*. 2013;35(2):121–126. doi:10.4103/0253-7176.116232
- 3) Lorenz L, Rüegger CM, O'Curraín E, Dawson JA, Thio M, Owen LS, Donath SM, Davis PG, Kamlin COF. Suction mask vs conventional mask ventilation in term and near-term infants in the delivery room: A randomized controlled trial. *J Pediatr*. 2018 Jul;198:181-186.
- 4) Kadam P, Bhalerao S. Sample size calculation. *Int J Ayurveda Res*. 2010;1(1):55–57.

III.10 The Rationale Behind Choosing the Appropriate Sample Size in Randomized Controlled Trials (Bill Rayburn)

While it may seem random at times, there is a science to selecting the appropriate number of patients for intended **Randomized Controlled Trials (RCT)**. Selecting too few patients could result in a false negative conclusion (type II error), whereas selecting too many leads to unnecessary expenditure of time and money. This section is intended to demystify the process of selecting an appropriate “n” for a given trial by going over the basic components of the calculation.

Components of Sample Size Calculations

While the final equations may differ based on the type of RCT and the intended outcomes, the same fundamental components exist throughout, which are listed below:

1. **Type I error (alpha):** As discussed in previous chapters of the EBM Guide, the designers of the study are required to set an alpha value for the data indicating their threshold for reaching a false positive. The alpha is most commonly set at 0.05, signifying that the researcher desires a <5% probability of drawing a false positive conclusion.
2. **Power:** On the other end of the spectrum, the researchers must determine the threshold for reaching a false negative conclusion or type II error (beta). As discussed again in Chapter 13, the calculation for the power of the study is $1 - \beta$. Conventionally, the beta is set at 0.20, which indicates a <20% probability of obtaining a false negative conclusion. Therefore, the typical power is $1 - 0.2$ or 0.8. As with alpha, the researcher can alter the power of the study to affect an individualized set of thresholds for type I and type II error. It should be noted that these values can (should) only be altered prior to data collection.
3. **The smallest effect of interest:** Otherwise known as the minimally clinically relevant difference (MCRD), the third component begins to introduce subjectivity into the sample size calculation. The MCRD is the difference that the investigator believes to be clinically and biologically possible. As one could imagine, a trial that anticipates large range of clinically significant values would require a smaller sample size than one investigating a smaller effect of interest. However, if the investigator or reader would not find a smaller effect to be clinically relevant, than the proposed trial would ultimately have limited impact. As it will be further described later, the sample size is related to the inverse square of the MCRD. Therefore, even small changes in MCRD can have a large impact on sample size. For example: if one would need 1000 subjects to detect an absolute difference of 4.8%, then 4000 subjects per treatment group would be required to detect a 2.4% difference.
4. **Variability:** The final component, sample size calculations are based on using the population variance of a given outcome variable. As with MCRD, the variability is typically an unknown quality that must be estimated by the investigators. Commonly, investigators will use an estimate based on a pilot study or information from a previously performed study. As it will be demonstrated below, variance is directly proportional to sample size.

Assembling the Components into a Sample Size Calculation

In this section, we will demonstrate a few proof-of-concept calculations using the elements described above. While the scope of every calculation for sample size is outside of the scope of this guide, an understanding of the relationship of the variables is invaluable in appreciating the decisions of the investigators and in designing future trials. As a note, the calculations do change based on the type of outcome measured and the type of trial (superiority, non-inferiority, etc.), highlighting the importance of utilizing biostatisticians. The simplest equation for sample size is the case of measuring a continuous outcome, which is listed below.

Box 1: Calculation for sample size in an RCT measuring a continuous outcome variable

<p>n = the sample size in each of the groups μ_1 = population mean in treatment Group 1 μ_2 = population mean in treatment Group 2 $\mu_1 - \mu_2$ = the difference the investigator wishes to detect σ^2 = population variance (SD) a = conventional multiplier for alpha = 0.05 b = conventional multiplier for power = 0.80</p> $\frac{n = 2 [(a + b)^2 \sigma^2]}{(\mu_1 - \mu_2)^2}$

Going back to our previous section, one can note that the final answer is 2 multiplied by the sum of the first two components squared times the square of the fourth component, all divided by the square of the third component. Unfortunately, it is important to keep in mind that the actual numbers used in the equation for “a” and “b” above are the **z-scores** for alpha and power, which can be through a quick internet search for a given alpha or power. For the sake of going through a sample calculation, the z-score for an alpha of 0.05 is 1.96, and the z-score for a power of 0.8 is 0.842.

Now, let us do a calculation for a theoretical trial investigating the effect of a new hypertensive drug on systolic blood pressure. The investigators determine that the minimal clinically relevant difference is 15mmHg for this case and based on past clinically trials determine that the variance is 20mmHg. Inputting all the variables, we have: $2[(1.96+0.842)^2 \times 20^2]/(15^2) = 27.9$ or 28 patients per group. Now, if the MCRD was determined to be 10mmHg, then the equation would be $2[(1.96+0.842)^2 \times 20^2]/(10^2) = 62.8$ or 63 patients per group. Small changes in any of the four components can lead to drastically different necessary sample sizes, highlighting the importance of accurate calculations when designing a study. As one might expect, there are multiple online calculators available to simply calculate the necessary sample size using the variables above, one of which will be included under the references section. For more information on the formulas for trial types not discussed here, please see the paper by Baoliang Zhong listed below.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

References:

1. Marlies Noordzij, Giovanni Tripepi, Friedo W Dekker, Carmine Zoccali, Michael W Tanck, Kitty J Jager, Sample size calculations: basic principles and common pitfalls, *Nephrology Dialysis Transplantation*, Volume 25, Issue 5, May 2010, Pages 1388–1393, <https://doi.org/10.1093/ndt/gfp732>
2. Zhong B. How to calculate sample size in randomized controlled trial?. *J Thorac Dis.* 2009;1(1):51-54. Online Clinical Calculator: <https://clincalc.com/stats/samplesize.aspx>

Submitted 6/2020

III.11 Blinding in Randomized Controlled Trials (Julia Harrison, GSM4)

Definition of Blinding:

The act of concealing or masking the true intervention of a randomized clinical trial from study participants and/or clinicians and data assessors to eliminate measurement bias (1).

Goal:

To prevent different treatment of groups in a trial and to prevent different interpretation or assessment of outcomes between groups. In other words, blinding serves to prevent both performance bias and ascertainment bias.

Importance:

In a systematic review of 250 RCTs identified from 33 meta-analyses, researchers observed a significant difference in the size of the estimated treatment effect between trials that reported “double-blinding” compared with those that did not ($p = 0.01$), with an overall odds ratio 17% larger in studies that did not report blinding (2). There is a wide variation in the application of blinding in RCTs, and this is an area that can be improved in many cases.

There are 5 groups that should be considered when assessing blinding in randomized controlled trials. These are: 1) participants, 2) clinicians, 3) data collectors, 4) outcome adjudicators and 5) data analysts. In some cases, the data collectors, outcome adjudicators and or data analysts may be the same person. If possible, trialists should blind all five groups of individuals involved in trials. Often, otherwise well-designed RCTs fail to blind data collectors, outcome adjudicators, and/or data analysts even when patients and providers are blinded. This introduces the possibility for bias. For example, in one otherwise well-designed RCT of cyclophosphamide and plasma exchange in patients with multiple sclerosis in which outcome adjudicators were not blinded, neither active treatment regimen was superior to placebo when assessed by blinded neurologists. However, there was an apparent benefit of treatment with cyclophosphamide, plasma exchange and prednisone when un-blinded neurologists performed the assessments (3).

Terminology in Randomized Controlled Trials:

Open Label- There is no blinding of any party in an open label trial.

This should only be conducted if blinding is deemed impossible or unethical, for example in comparing a medical versus a surgical intervention in which you cannot conceal the treatment from the surgeon and patients, although even in this scenario there are other components of the trial that can be blinded, such as data collection, outcome adjudication, and data analysis. Be skeptical when approaching the results of an open label trial.

Single Blind- the nature of the intervention is concealed from the participants.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

If participants are not blinded, knowledge of group assignment may affect their behavior in the trial and their responses to subjective outcome measures. This increases non-adherence and attrition among participants in the control group who are aware they are not receiving active treatment and may make it more likely that they seek additional treatment or leave the trial.

Double Blind- the nature of the intervention is concealed from both the participants and the research team.

If clinicians are not blinded, they are much more likely to transfer their attitudes to participants or to provide differential treatment to the active and placebo groups

Triple Blind- the nature of the intervention is concealed from participants, the research team, and the data collectors, outcome adjudicators, and data analysts (may be the same person).

Crucial to ensuring unbiased ascertainment of outcomes. Most important in subjective outcomes. However, seemingly objective outcomes often require some degree of subjectivity and therefore are at risk of bias as well.

** NOTE- the terms double-blinded and triple-blinded are often ambiguous and used inconsistently- it is preferable that studies disclose the exact groups that were blinded.

Techniques for blinding:

Medical trials- placebo medication with same appearance, smell, taste as trial medication

Surgical trials- more difficult to achieve total blinding. However, there are techniques for partial blinding that are underutilized, such as using an independent individual unaware of the treatment allocation for data collection and analysis, concealing incisions or scars from patients and providers for several days post-op, and digitally altering radiographs to mask the type of implant in orthopedic procedures .

References:

- 1) Guyatt, G., Rennie, D., Meade, M. O., & Cook, D. J. (Eds.). (2008). *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice* (2nd ed.). New York: McGraw-Hill
- 2) Schulz KF, Chalmers I, Hayes RJ, et al. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273:408–12.
- 3) Noseworthy JH, Ebers GC, Vandervoort MK, et al. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology*. 1994;44:16–20.
- 4) Karanicolas PJ, Farrokhyar F, Bhandari M. Practical tips for surgical research: blinding: who, what, when, why, how?. *Can J Surg*. 2010;53(5):345–348.

III.12 Randomization (Chris Del Prete)

Randomization represents the statistical cornerstone of modern-day clinical trials. Since at least the early 1990s, most reputable academic journals will not publish a clinical trial that does not feature randomization, except in rare cases in which it is not appropriate. While the value of randomization is universally accepted, there is considerable confusion even among researchers regarding the different types of randomization and how these are applied in practice. This section will review the basics of randomization and discuss in depth different randomization techniques with a particular focus on techniques used in the biomedical literature

Basic definition and importance

Statistical randomization in the context of clinical trials refers to the process of randomly assigning patients to either the treatment group or the control group. In other words, it is a method of treatment allocation. The importance of randomization rests on the premise that in any given study population, there will be multiple baseline characteristics that could potentially act as confounders when assessing whether a true difference exists between treatment and control groups.

Randomly assigning patients to a treatment group or control group gives researchers the best chance of equally distributing both known and unknown cofounders between study groups. In the event that researchers find a significant intergroup difference, if subjects have been randomly assigned there is a greater likelihood that this is a true treatment effect, i.e., one not due to confounding variables. Whether randomization successfully distributed patients into groups with similar baseline characteristics can be assessed in a table that accompanies most RCTs. Another advantage of randomization is that—provided it is performed in a concealed manner—selection bias is eliminated. The researcher cannot select which patients will receive an intervention and which a control.

Types of Randomization and Method

Simple Randomization

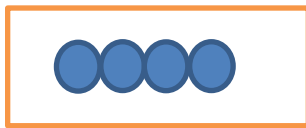
Think of this method as a coin toss. For example, if assigning patients to a treatment group vs a control group, heads = treatment, tails = control. Instead of coins, researchers now use random numbers generated by a computer or random number tables found in statistics textbooks or online. A researcher, ideally one either offsite or otherwise independent of the research being conducted, first sets parameters for how the random numbers generated from a table or computer will be used. For example, we can decide to assign patients who end up with even numbers to the treatment group and odd numbers to the control group. The researcher then decides where in the chart to start and which direction to proceed in if using a table (up, down, diagonally).

This type of randomization works best in larger clinical trials, those with at least 100-200 patients. In trials with smaller numbers of patients (especially those less than about 50 patients), this method of randomization can result in uneven distribution of numbers of patients between groups by chance alone.

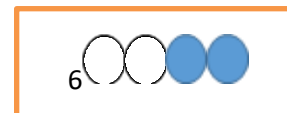
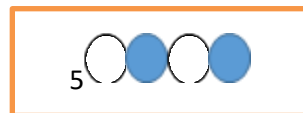
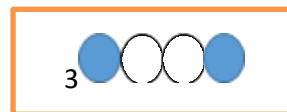
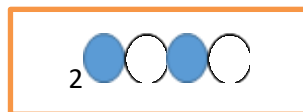
Block or restricted randomization

Sometimes, particularly with small numbers of patients, it is preferable to randomize patients in such a way that equal group sizes result. When this is the intent, a process called block or restricted randomization is used. A block refers to a set number of patients as designated by the researcher. Once this number or block size is set and patients have been assigned to a block, patients within each block are then assigned to either the treatment or control groups. Each block is balanced: that is, each block has an equal number of patients in treatment or control groups. Next, every possible balanced combination is created within each block. Then, blocks are randomly selected until every participant is assigned to a group. Below is a visualization of this process for a theoretical trial of 40 patients:

- 1) Assignment of block size. In this case, the researcher has selected 4 patients per block.



- 2) Possible balanced combinations (dark circles= treatment, clear = control)



- 3) Random selection of the blocks above to assign patients to treatment or control (10 blocks of 4 patients each = 40 patients). These are the patterns we will apply until all 40 patients have been assigned:



We have now ensured equal numbers of patients in control and treatment groups. (Adapted from Kang et al and Altman et al).

Stratified Randomization

Ever wonder how researchers achieve that balance of baseline characteristics you see in a typical table 1? Stratified randomization is one method utilized to achieve this balance.

Essentially, this process creates the blocks above, but unlike block randomization alone, the blocks are created separately for each baseline characteristic or stratum that needs to be controlled for, like age, sex or functional status. Patients are then assigned to each block as above. This results in equal numbers of patients in each stratum assigned to treatment or control groups.

Example (adapted from Kang et al): Say you have a treatment and control group involving a study that wants to control for two different covariates, sex, and BMI. So, you want your treatment and control groups to have equal numbers of male/female patients and BMI categories (underweight, normal, overweight).

So, what you do first is come up with all possible block combinations. In this case it's 6: you can have someone be a 1. male underweight, 2. male normal, 3. male overweight, 4. a female underweight, 5. female normal, 6. female overweight. You then do simple randomization (i.e., a coin flip) to assign participants within each block to treatment or control. So now, you assign your male underweight patients to treatment or control with a coin flip. Then assign your female overweight patient to treatment or control, and so on. On average this will result in balanced groups because it relies on a coin toss.

Covariate adaptive randomization or minimization

The final method of randomization we will address here is covariate adaptive randomization, sometimes referred to simply as “minimization.” I prefer the latter term because it adequately describes the intent of the process: namely, we are attempting to minimize imbalances in covariates/baseline characteristics by randomizing as above, but in a smarter fashion.

There are a number of specific ways this can be accomplished, and each is named after a famous statistician. I will detail the Taves method below.

As before, the researcher identifies important covariates that require balancing between treatment and control groups. As each patient is recruited into the study, a decision is made (by an independent/unaffiliated researcher or computer) on which group to place the patient in based on the balance of covariates already included in the study. In other words, the patient will be placed into the group (treatment or control) that *minimizes the imbalance* in treatment or control groups. This can be done for multiple variables of interest simultaneously. Ultimately, the advantage of sorting patients this way is that it allows for the continuous and balanced recruitment of individuals into a trial.

Conclusion

Proper randomization is an essential step in the process of developing a rigorous randomized controlled trial. It allows for control of baseline characteristics and other patient specific factors that could otherwise confound analysis. It also eliminates selection bias and provides the statistical backing for assessments of correlation and efficacy. There are a variety of statistical techniques used to randomize patients into trials, though covariate adaptive randomization or minimization is the most practical for a modern-day RCT.

References:

1. Altman, D, Bland J, "Treatment Allocation in Controlled Trials: Why randomize?" BMJ 1999 (318): 1209 Altman, D, Bland J, "How to randomize" BMJ 1999 (319): 703-704.
2. Kendall, J. "Designing a research project: randomized controlled trials and their principles." Emergency Medicine Journal 2003 (20): 164-168
3. Kang M, Ragan B, et al. "Issues in Outcomes Research: An overview of Randomization techniques in Clinical Trials." Journal of Athletic Training 2009(43): 215-221.

October 2017

III.13 Intention to Treat vs. As-Treated Analysis (Jen Frampton and Resham Ramkissoon)

- Intention to treat analysis is a method of analysis for randomized controlled trials in which all patients randomly assigned to one of the treatments are analyzed together, regardless of whether or not they completed or received that treatment, in order to preserve randomization.
- In randomized controlled trials – data will be calculated from the intention to treat analysis. In some cases, they will compare this with the as-treated population.
- The as treated group is the actual treatment received – not what they were supposed to receive.
- The as treated group is more accurate results from the study, although introduces potential bias because of the deviation from the original intention groups.
- If the results are similar, then they can be helpful in interpreting the study and the study's results.

A focus on the Intention to Treat (ITT) analysis

Two major problems encountered with randomized control trials (RCTs) are patient non-compliance and missing outcomes. An intention-to-treat (ITT) analysis offers a solution to this problem; every subject that was randomized into different study groups are included in the analysis. ITT disregards non-compliance, deviations in protocol, subject withdrawal, et cetera. An ITT analysis can be described as “once randomized, always analyzed”.

When there are changes to the study population after randomization (due to non-compliance, withdrawal, et cetera), there can be an overestimation of the effectiveness of an intervention or likelihood of a measured outcome. Adopting an ITT analysis avoids these overestimations and reduces biases. Other advantages of using an ITT analysis is preservation of a study's statistical power by maintaining the original sample size, limits “ad hoc” subgroups in a study population, and minimizes type I errors (false positive results).

A major criticism of an ITT analysis is the likelihood of encountering type II errors (false negative result). For example, a patient, who did not receive treatment, which is included in the treatment arm of a study will detract from the true efficacy of the treatment. Primary and secondary outcomes may differ greatly between noncompliant, dropouts, and compliant subjects. This will make interpretation difficult if there are large numbers of subjects that “cross-over” between treatment arms.

An ITT analysis can be applied effectively if care is taken to minimize missing responses and there is continued follow-up for subjects who withdraw from treatment. Often enough, missing data points are hard to avoid and can be addressed by the “Last Observation Carried Forward” (LCOF) method; the last available measurement for individuals at the time point prior to withdrawal is retained in the analysis.

Furthermore, it is often helpful to look at the study outcomes/results from two perspectives; the ITT analysis and the “Per Protocol” (PP) analysis (which excludes study protocol violators). The US Food and Drug Administration (FDA) adopts a practice of looking at outcomes from both the ITT analysis and reduced subset in a PP analysis. If there are differences between the outcomes in these two types of analysis; this will need to be reconciled by the study investigators as a type 1 error, a type 2 error, or if the study was conducted poorly.

September 2017

III.14 Primary vs Secondary Outcomes, Dichotomizing, and Selection of Endpoints (Kayla Hatchell, GSM4)

Primary outcomes are the outcomes that investigators consider to be the most important in the study. Ideally, they are a clinically relevant event that is significant for the patient and are directly related to the primary goal of the trial. They should be the outcomes that are best at determining if the trial was a success or a failure and should be available in the clinical trial registry. A redefinition of the primary outcome after un-blinding the results is almost always unacceptable. Primary outcomes must be defined before a study is begun for two reasons:

- To reduce the risk of false-positive errors, resulting from testing many outcomes. With testing multiple outcomes, even if each outcome has a <5% chance of giving a false result, the more outcomes you use the higher chance that an outcome will have a false-positive result (Type I error).
- To reduce the risk of a false-negative error, by providing the basis for the estimation of sample size necessary for an adequately powered study (Type II error).

Secondary outcomes are other outcomes that are deemed important by the investigators. As stated above, identifying many secondary outcomes increases the probability that at least one secondary outcome will be a false positive. Secondary outcomes are not powered at the same level as the primary outcome as the sample size calculation is based on the primary outcome. If the power is too low for a secondary outcome, there may be a false-negative result. This it is best to pay the most attention to primary outcomes and interpret the secondary outcomes with more caution. Secondary outcomes can be important for forming hypotheses for future studies and aiding the interpretation of the primary outcome.

Categorical variables (defined in Chapter 29 of this guide) can be useful in labeling individuals as having or not having an attribute, such as hypertension, obesity, or high cholesterol.

Continuous variables have the advantage of requiring a smaller sample size to identify statistically significant differences.

One benefit of using continuous outcomes is the improved ability to achieve a greater study power. For example, a review of 76 orthopedic RCTs with sample sizes of < 50 patients in 2001 found that studies that reported continuous outcomes had a significantly greater study power than studies that reported dichotomous outcomes (2 categories) ($p=0.042$). Twice as many studies that reported continuous outcomes achieved conventionally acceptable study power (80% or more) than those that reported dichotomous outcomes ($p=0.04$).

In creating categorical variables, sometimes continuous variables, such as HbA1c, are converted into categories. For example, a cut-off of HbA1c = 6.5% can be used to distinguish a diabetic patient from a non-diabetic patient in a trial rather than reporting the values of HbA1c of participants. This process of converting continuous data into two groups is called **dichotomizing**. There can be downsides to this approach. For example, individuals close to but on opposite sides of the cut-off are characterized as being very different rather than very similar. Choosing a cut-off at the median of the study population is especially risky, as it can underestimate the extent of variation in outcomes between groups. It also makes meta-analyses difficult to conduct as results cannot easily be compared. Performing several analyses and choosing the “optimal” cut-off that achieves the minimum P value can overestimate the differences between the groups and can give a falsely low confidence interval.

Outcomes, or endpoints are described as “hard” or “soft”. A **hard endpoint** is well-defined and can be measured objectively. See examples of hard endpoints and considerations of each below:

All-cause mortality

- Considered the most unbiased endpoint: easy to measure, not readily subject to observer bias
- Represents an important event for the patient
- Downsides include that this endpoint generally requires a large sample size, depending on the risk of the patient population. Choosing to enroll only high-risk patients would decrease the necessary sample size but would make the generalizability of the results lower.

Adjusted all-cause mortality

- Uses survival regression models to adjust for any imbalance in prognostic variables that could be present between study groups
- Best to choose pre-specified clinically relevant prognostic factors (e.g., ejection fraction in CHF)

Cause-specific mortality

- Definitions of causes can differ between studies
- The number of events in each category is reduced, which reduces the statistical power of the analysis to detect any difference between the treatment groups

Rate of (re-)hospitalization

- There may be different thresholds for hospitalization in different centers and regions
- It may be difficult to define the main cause for hospitalization, especially in patients with several concomitant diseases

Composite variables

- Integrate or combine multiple variables into a single or composite variable by using a pre-defined algorithm

Soft endpoints are subjective measures that are considered clinically relevant. They are sometimes best to evaluate the effect of a treatment in earlier stages of disease when the hard endpoints above may be too infrequent.

Functional status

- May underestimate functional disability as there can be an apparent improvement when the patient merely reduces stressful activities
- An improvement is often evident in the placebo arm of trials

Quality of life or health-related quality of life

- Most common domains assessed are physical functioning, emotional/psychological well-being, social functioning, role functioning (including employment), disease-specific symptoms, and general health perceptions.
- Usually based on validated questionnaires administered before and after an intervention. Thus, the data cannot be measured as unrelated comparisons.
- Instruments may detect impairments but not provide information about causes.

Worsening symptoms

- May have imprecise definitions

Surrogate endpoints are laboratory measurements or physical signs that are used as substitutes for clinically meaningful endpoints that measure directly how a patient feels, functions, or survives. Changes induced by a therapy on a surrogate endpoint are expected to reflect changes in a clinically meaningful endpoint. Surrogate endpoints must have biological relevance, and the effect of the treatment on the surrogate must be able to predict the effect of the treatment on the clinical outcome. An advantage of using them is that usually only a few hundred patients are required to achieve a high statistical power.

References:

1. Andrade, C. The primary outcome measure and its importance in clinical trials. *J Clin Psychiatry* 2015;76(10):e1320-e1323.
2. Altman, DG and Royston, P. The cost of dichotomizing continuous variables. *BMJ* 2006 May 6; 332(7549): 1080.
3. Mark, D. Assessing quality-of-life outcomes in cardiovascular clinical research. *Nature reviews cardiology* 2016;13(5):286-308.
4. Temple RJ. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT, editors. *Clinical measurements in drug evaluation*. New York: J. Wiley, 1995.
5. Zanolla, Luisa and Zardini, Piero. Selection of endpoints for heart failure clinical trials. *The European Journal of Heart Failure* 2003;(5):717-723.
6. Bhandari, M, Lochner H, Tornetta P. Effect of continuous versus dichotomous outcome variables on study power when sample sizes of orthopedic randomized trials are small. *Arch Orthop Trauma Surg.* 2002;122(2):96-8.

III.15 Clinical Outcome Assessments and nuances of trial design in Neurology (Dennis Obat, GSM4)

Introduction:

Making a "PICO" question (population, intervention, control/cointervention, outcome) is a great framework to have when it comes to designing and analyzing clinical trials. The selection of the study population and interventions are intuitively important, but a study can fall flat without a robust and relevant method of measuring the outcome.

When I started the EBM course, I thought the most difficult part of the course would be wrapping my head around the different mechanisms of, or reasoning behind, the interventions investigators used in their studies. However, I found that I spent most of my time trying to understand the different rating scales used to measure treatment response and figuring out whether these captured what the investigators had intended them to reflect.

The major outcome measures I commonly encountered in clinical trials were mortality/survival, biomarkers or surrogate outcomes, and clinical outcome assessment (COAs). All-cause mortality or

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

survival is a direct and obvious measurement and are dichotomous outcomes; biomarker assessment is just as easy to understand (example: “does drug X reduce the expression of pathogenic protein Y in patients with disease Z compared to placebo?”). COAs required a more nuanced understanding. In this essay, our focus will be on COAs; paying particular attention to the modified Rankin Scale (mRS) which was the one I encountered the most over the course of this elective.

Types of outcome measures	Example
Direct/objective	All-cause mortality, 5-year survival
Biomarker	Protein expression, blood pressure
Clinical outcome assessment (COA) a. Clinician-reported b. Patient-reported c. Observer-reported d. Performance outcomes	mRS QOLIE-31, PHQ-9 ALS-FRS

Table 1: Types of outcome measures¹. mRS: modified Rankin scale; QOLIE-31: quality of life in epilepsy. PHQ-9: patient health questionnaire. ALS-FRS: amyotrophic lateral sclerosis functional rating scale

What is a COA and what are the different types?

The FDA defines a COA as “measure that describes or reflects how a patient feels, functions, or survives”¹. The Japan Supportive, Palliative and Psychosocial Oncology Group (J-SUPPORT) expands on this by describing COA as “any assessment that may be influenced by human choices, judgment, or motivation and may support either direct or indirect evidence of treatment benefit”².

There are different categories of COAs (Table 1). However, it is worth noting that some COAs may fit into more than one category. A clinician-reported COA is a measure obtained by a clinician, or other trained healthcare professional, after observing a patient. These measures involve clinical judgement or interpretation of observable signs and symptoms thought to be related to a patient’s condition. A good example of this is the NIH Stroke Scale (NIHSS).

Patient-reported outcomes, on the other hand, are measures based on a patient’s report of their condition and do not involve any amendment or interpretation by a clinician or anyone else for that matter and are recorded as such (examples: a patient saying their pain intensity is a 7 on a scale of 1 to 10, a seizure diary).

Observer-reported measures are those taken by someone other than the patient or a healthcare professional. These are usually reported by someone who is familiar with the patient and is able to observe them regularly (example: a parent’s report of the number of times their infant vomited). Like patient-reported outcomes, these are recorded as given and involve no amendment or interpretation by a clinician.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Performance outcome measures are used to evaluate a patient's performance in a task when instructed by a healthcare professional. An example of this would be the Timed Get-up and Go test or the MOCA test used to assess cognition. These usually require patient cooperation and motivation.

When designing a study, it is important that investigators establish several factors regarding their chosen COA. Some of these factors include: Does the COA measure what it is supposed to? Is it reliable? And is the outcome measured clinically relevant/important to patients? Most of the studies I did during my EBM elective were looking at outcomes of patients after a stroke. For this reason, I will discuss COAs in the context of the modified Rankin Scale (mRS) in the rest of this essay.

What is the mRS?

The modified Rankin scale (mRS) is a clinician-reported COA used to document the degree of disability or dependence when it comes to performing activities of daily living. The original scale was developed by Dr. John Rankin in 1957 to document functional recovery after a cerebrovascular accident in patients >60 years old at the time of discharge. It has become the most widely used COA for stroke clinical trials³. The original scale consisted of five categories but two additional categories were added. In the mRS, grade 1 of the original RS is replaced by grade 0 and 1 which allows for a finer discrimination of mild strokes. Grade 6 allowed one to denote mortality in the mRS (Table 2).

Score	Description	
	Original RS	mRS
0	n/a	No symptoms at all
1	No significant disability: able to carry out all usual duties	No significant disability: despite symptoms, able to carry out all usual duties and activities
2	Slight disability: unable to carry out some of previous activities but able to look after own affairs without assistance	Slight disability: unable to carry out some of previous activities but able to look after own affairs without assistance
3	Moderate disability: requiring some help but able to walk without assistance	Moderate disability: requiring some help but able to walk without assistance
4	Moderate disability: unable to walk without assistance and unable to attend to own bodily needs without assistance	Moderate disability: unable to walk without assistance and unable to attend to own bodily needs without assistance
5	Severe disability: bedridden, incontinent, and requiring constant nursing care and attention	Severe disability: bedridden, incontinent, and requiring constant nursing care and attention
6	n/a	Death

Table 2^{3,4}: comparison between the original RS and the mRS

Validity

Validity is the extent to which a test measures what it is intended to measure. A good COA should accurately measure the outcome it claims to measure. One can assess validity of a COA by using

surrogate tests of the same outcome and comparing results from those tests to the COA under evaluation.

Take the example of outcome X that is measured by multiple COA tests A, B, and C. If I wanted to make or use a new COA test (D), D would be considered valid if a patient with a “bad outcome” on test D also has a “bad outcome” when we measure X using A, B, and/or C. This convergent criterion is an important property of valid tests.

Now, the mRS was designed to measure a patient’s recovery from a stroke. It follows, therefore that a patient with a more severe stroke (by location, type, or volume etc.) should be more “disabled” as reflected by a higher mRS. This has been confirmed by multiple studies that show patients with more severe strokes have poorer outcomes based on the mRS. For example, a study evaluating this found that higher baseline NIHSS scores were associated with worse outcomes (mRS > 2) at 3 months ($p < 0.00001$)⁵. Another showed that the size of a stroke lesion on imaging is correlated to higher RS scores ($p < 0.001$)⁶. These studies proved the relationship between stroke severity and mRS. And this was emphasized when other studies showed that poorer outcomes on mRS also correlate to poorer outcomes on other tests of disability such as the Barthel Index which looks at a patient’s ability to perform activities of daily living (ADLs)^{7,8}. Based on these studies, the mRS passes the validity test for assessing functional outcomes after a stroke.

Reliability

Reliability is the extent to which a COA consistently reproduces the results of an outcome it claims to measure. The mRS has been shown to have a high test-retest reliability⁹. The test also has a near-perfect inter-rater reliability even when administered in a different language^{9,10}. This high inter-rater reliability is in part due to the development of a standardized questionnaire (Figure 1). One can imagine how difficult it would be to compare stroke trials if the results varied significantly based on factors besides stroke severity/time from stroke.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Q1: Do you have any symptoms that are bothering you?	<input type="radio"/> No	<input checked="" type="radio"/> Yes
Q2: Are you able to do the same work as before?	<input checked="" type="radio"/> No	<input type="radio"/> Yes
Q3: Are you able to keep up with your hobbies?	<input checked="" type="radio"/> No	<input type="radio"/> Yes
Q4: Have you maintained your ties to friends and family?	<input checked="" type="radio"/> No	<input type="radio"/> Yes
Q5: Do you need help making a simple meal, doing household chores, or balancing a checkbook?	<input checked="" type="radio"/> No	<input type="radio"/> Yes
Q6: Do you need help with shopping or traveling close to home?	<input checked="" type="radio"/> No	<input type="radio"/> Yes
Q7: Do you need another person to help you walk?	<input checked="" type="radio"/> No	<input type="radio"/> Yes
Q8: Do you need help with eating, going to the toilet, or bathing?	<input checked="" type="radio"/> No	<input type="radio"/> Yes
Q9: Do you stay in bed most of the day and need constant nursing care?	<input checked="" type="radio"/> No	<input type="radio"/> Yes

mRS = 2
modified Rankin Scale

Slight disability
Able to look after own affairs without assistance, but unable to carry out all previous activities.

Figure 1. Sample mRS questionnaire from MDCalc (<https://www.mdcalc.com/calc/10430/modified-rankin-score-9q-mrs>)

Limitations

The mRS is not without its flaws. Comorbidities such as cardiovascular disease, diabetes, and depression impact the recovery of patients post-stroke and can have a direct effect on the mRS¹¹. Other disability scales such as the Barthel Index for ADL share this limitation.

Because the process of obtaining the data relies on interviewing patients, COAs are susceptible to variations that may be unrelated to the patient's true clinical status. A variation in the motivation of a patient can cause differences between successive administrations of a test or between patients. The judgement of the rater/interviewer may also lead to different ratings based on prior experience or biases.

How is this clinically relevant?

This is, perhaps, the most difficult property of COAs to determine. Is the difference between an mRS of 5 and 4 the same (read: as clinically relevant) as the difference between 2 and 1 or 1 and 0? There is limited data on how well the mRS captures the effect of an intervention and where an investigator should draw the line to determine whether their intervention was beneficial or not. The use of non-dichotomous scores on a continuous scale of values is a persistent problem for these measures.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

To address this, many have proposed clustering/dichotomizing the mRS scores once all the data have been collected. An example of this is when a study groups all mRS scores of 0-1 as “excellent outcome” and denotes mRS>3 as “poor outcome” when doing the statistical analysis. The optimal point of dichotomization is affected by the anticipated distribution of mRS based on initial stroke severity. This distribution determines the level of the scale that a treatment effect is most likely to be observed¹². Unfortunately, for investigators, there is no way of knowing this ahead of time. That said, dichotomizing the scale makes it easier for an investigator to detect the benefit, or lack thereof, of an intervention than it would have been if one looked at “mean change from baseline mRS”.

However, this cannot be done for all COAs. And for those COAs that cannot be clustered or dichotomized to tease apart the differences between groups, the question of when a statistically significant change translates to a clinically significant change will continue to persist.

Final thoughts

Understanding the attributes of and evaluating the evidence behind a COA will help investigators construct more robust studies to evaluate the effect of an intervention. Clinicians, too, should be aware of the characteristics of a COA and interpret the results of any study within that context to avoid misapplying trial data to their patients.

Digging deeper into COAs (and the EBM course at large) has given me a space to more effectively appraise the medical literature. I have learned skills that will stand me in good stead over the course of my career. One question that kept coming up over the course of this elective was “is this clinically important?” This question serves as a reminder for me that practicing EBM does not preclude using one’s clinical judgement and, most importantly, listening to the patient and figuring out what is meaningful *to them*—something that many COAs may fail to do. There will never be a perfect COA, but practicing individualized patient-centered EBM is a good place to start when applying trial data to our patients.

References

1. Center for Drug Evaluation and Research. *Clinical outcome assessment(Coa):faqs*. U.S. Food and Drug Administration. FDA. Available at <https://www.fda.gov/about-fda/clinical-outcome-assessment-coa-frequently-asked-questions#COADefinition>. (Accessed: January 21 2023)
2. Clinical outcome assessment. *J-SUPPORT*. Available at <https://www.j-support.org/en/rating/index.html>. (Accessed January 21 2023)
3. Wilson JT, Hareendran A, Grant M, Baird T, Schulz UG, Muir KW, Bone I. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

modified Rankin Scale. *Stroke*. 2002 Sep;33(9):2243-6. doi: 10.1161/01.str.0000027437.22450.bd. PMID: 12215594.

4. Rankin J. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. *Scott Med J*. 1957 May;2(5):200-15. doi: 10.1177/003693305700200504. PMID: 13432835.
5. Nedeltchev K, der Maur TA, Georgiadis D, Arnold M, Caso V, Mattle HP, Schroth G, Remonda L, Sturzenegger M, Fischer U, Baumgartner RW. Ischaemic stroke in young adults: predictors of outcome and recurrence. *J Neurol Neurosurg Psychiatry*. 2005 Feb;76(2):191-5. doi: 10.1136/jnnp.2004.040543. PMID: 15654030; PMCID: PMC1739502.
6. Kluytmans M, van Everdingen KJ, Kappelle LJ, Ramos LM, Viergever MA, van der Grond J. Prognostic value of perfusion- and diffusion-weighted MR imaging in first 3 days of stroke. *Eur Radiol*. 2000;10(9):1434-41. doi: 10.1007/s003300000501. PMID: 10997432.
7. Wolfe CD, Taub NA, Woodrow EJ, Burney PG. Assessment of scales of disability and handicap for stroke patients. *Stroke*. 1991 Oct;22(10):1242-4. doi: 10.1161/01.str.22.10.1242. PMID: 1833860.
8. Lai SM, Duncan PW. Stroke recovery profile and the Modified Rankin assessment. *Neuroepidemiology*. 2001 Feb;20(1):26-30. doi: 10.1159/000054754. PMID: 11174042.
9. Wilson JT, Hareendran A, Hendry A, Potter J, Bone I, Muir KW. Reliability of the modified Rankin Scale across multiple raters: benefits of a structured interview. *Stroke*. 2005; 36: 777–781.
10. Van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van GJ. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke*. 1988; 19: 604–607
11. Nichols-Larsen DS, Clark PC, Zeriongue A, Greenspan A, Blanton. Factors influencing stroke survivors' quality of life during subacute recovery. *Stroke*. 2005; 36: 1480–1484.
12. Broderick JP, Adeoye O, Elm J. Evolution of the Modified Rankin Scale and Its Use in Future Stroke Trials. *Stroke*. 2017 Jul;48(7):2007-2012. doi: 10.1161/STROKEAHA.117.017866. Epub 2017 Jun 16. PMID: 28626052; PMCID: PMC5552200.

Submitted 1-23-2023

III.16 Understanding the Continuing Conundrum of Continuous Variables (Swathi Krishnan, GSM4)

In the hundreds of randomized controlled trials that are published every year, a significant number of them use continuous outcome variables as the outcome, exposures, or covariates of the study. A continuous variable, as opposed to a categorical variable, is one that can hypothetically take on any value in a given interval. These measures can range from quality-of-life measures, like pain scales, to age, or blood pressure. The majority of studies that use continuous variables to report an intervention's effect on an outcome as compared to a control group use the difference in means.

Although continuous variables are common in clinical studies, analyzing the results in terms of clinical relevance and applicability can be challenging. Unlike with the categorical variable, there is no easy way to calculate a 'number needed to treat' for continuous variables from a difference in means, and so it is difficult to interpret the results that can then be translated to clinical practice and more precise patient communication. Another shortcoming of a difference in means is that this type of measure is an average amongst a group of individuals; there is no way of understanding which proportion of people might achieve a given degree of benefit or harm. Meta-analyses of studies with continuous variables also suffer from trying to compare different scales from trials that are essentially studying the same measure.

One solution that researchers and analysts have come up with to deal with some of these problems categorizing or dichotomizing the continuous measure- this means grouping values in two categories. Usually, the researchers would create a set cutoff-value, above which would be one category and below which would be the other. Although many researchers and statisticians have come up with various schema for categorizing continuous variables, some of the more popular methods include Cox & Snell (1989), Suissa (1991), Hasselbad and Hedges (1995), Furukawa (1999), and Kraemer & Kupfner (2006) to name a few. While this discussion is not meant to delve deeply into the statistical methods of these various groups, it is interesting to see how prevalent the conundrum of the continuous variable is, and how creating a binary system for analysis is an ongoing and evolving process. As a general summary, conversion methods that facilitate the translation of continuous measures into clinically relevant estimates of effect such as relative and absolute risk or numbers needed to treat in meta-analyses rely on assumptions of data distribution, control group responses and sample size differences between the studies being compared. Importantly, a firm understanding of baseline risk is required to interpret the degree of benefit from a difference in population means.

It is interesting to note that this conversion of continuous measures to a binary model is a mindset that physicians instinctually use in clinical practice- hypertensive vs non-hypertensive, dyslipidemic vs non-dyslipidemic, diabetic vs. non-diabetic, etc. All these categories take either the blood pressure, cholesterol level, or HgbA1c— which are all continuous variables— and are transformed into categorical variables for the purposes of treating patients. This mindset is not necessarily a bad thing, given that clinical decision making often is a binary system— treat vs not treat, cancerous vs benign, normal vs abnormal etc.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

While the clinical setting has shown the sensibility of using dichotomous decision points, there can be dangers in utilizing such strategies in a research setting. At the most basic level, there is a danger in loss of both information and power, as in essence, the process of categorization is an extreme form of rounding. Prior analysis has also shown that categorized continuous variables may require more than a 40% increase in patients in order to achieve the same power as using continuous variables. In smaller, single studies, this loss of power is even more impactful. Another problem with splitting individuals into only two categories, is that the inter-group variation of individuals becomes marked. Also, individuals on either side of the cutoff point, who would generally be shown to have similar outcomes in reality may have very different outcomes. This highlights the challenges of selecting a cutoff point to begin with. Depending on what value is used to designate normal vs abnormal, significant associations with the given outcome can change. In fact, there are certain studies looking at age in relation to atrial fibrillation stroke risk, which have shown that compared to the binary categorization model or even a 3-category model, expressing age as a cubic polynomial best explained the data in both men and women. This brings into question whether other modeling strategies (natural splines, fractional polynomials, or non-parametric techniques) might have better results; on the flip side, with ever more complicated statistical research models, clinical applicability becomes less clear.

Overall, it is important to recognize the value of the increasing number of RCTs published that use continuous variables as the primary outcome measure, without disregarding such studies as less valid or having less statistical integrity. The use of continuous measures is generally thought to have more power (by categorizing data you have fewer degrees of freedom).

The ability to categorize continuous variables can be a useful way to report and interpret continuous measures, though one should be cautious of the potential pitfalls as noted above. These methods are merely one more tool that, if used in the proper clinical setting, can aid in evidence based medicine and shared decision making.

References:

1. van Walraven C, Hart RG. Leave 'em alone - why continuous variables should be analyzed as such. *Neuroepidemiology*. 2008;30(3):138-9. doi: 10.1159/000126908. Epub 2008 Apr 17.
2. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *BMJ*. 1998 Feb 28;316(7132):690-3. Review
3. Mayer M. Continuous outcome measures: conundrums and conversions contributing to clinical application. *BMJ Evid Based Med*. 2019 Aug;24(4):133-136.
4. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006 Jan 15;25(1):127-41.

III.17 Confounding and how to mitigate its impact (Ashley Dunkle, GSM4)

Introduction – Confounding and the Importance of Validity

The purpose of research studies in medicine and epidemiology is to discover if there is an association between an exposure and an outcome. If there is an association and the study is valid, we can make a causal inference that an exposure *causes* an outcome. By doing so, we can then intervene on the exposure in order to prevent (or enhance if it is a positive one) an outcome.

One of the most important factors when assessing a study that claims an association (or lack thereof) is to determine validity. If a study is valid, we can assume the observed effect of the exposure on the outcome is true, or rather, there is a very small chance the observed association is due to random chance. If an association is observed, it may be that it is true, and we have learned something about our environment and physiology that can help to improve patient and public health. However, there are three alternative explanations for why an association is or is not observed in a study due to lack of validity. (1) Random error, (2) Bias, or (3) Confounding.

(1) **Random error** is the probability that the observed association is due to random chance. This is where the p-value and confidence intervals come in. Random error is always a possibility in studies, just as the role of a dice can get different results each time. However, we reduce random error by increasing the study sample size and powering a study appropriately, so we are more likely to observe the true effect if one exists.

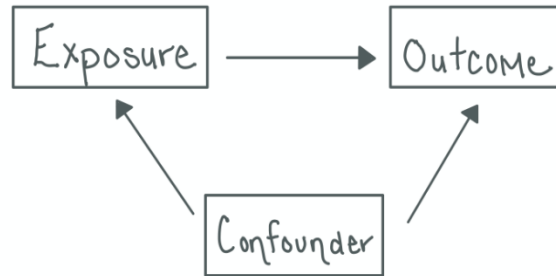
(2) **Bias** is systematic error due to study design or researcher actions throughout the course of the study. Systematic error will lead to the same erroneous results each time, unlike random error. There are excellent explanations and examples of bias in this study guide, so check them out!

(3) **Confounding**...can be confounding. This chapter aims to dive into what confounding is so you can understand how to keep an eye out for it in studies, make valid conclusions from what you are learning, and how you can prevent it in your own research.

What is confounding?

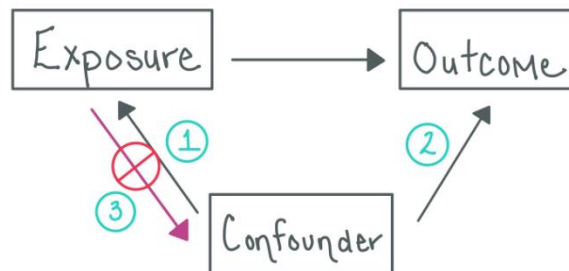
In very simple terms, confounding is the **mixing of effects** between an exposure and an outcome by a third variable known as a confounder. When a confounder is present, the association between an exposure and an outcome is distorted because there is a relationship between the exposure and the confounder and the outcome and the confounder. However, pay attention to the direction of the arrows in the description of these associations.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine



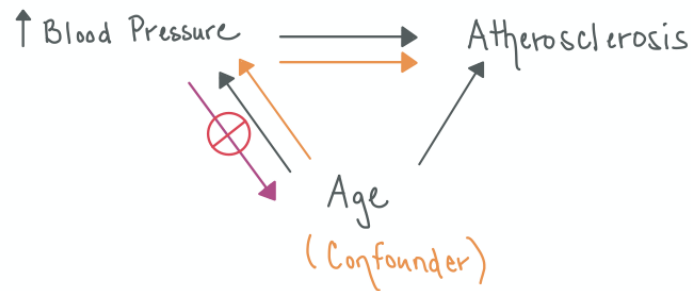
Three things must be true for a variable to be a confounder:

1. A confounder must be associated with the disease (either as a cause or as a proxy for a cause but not as an effect of the disease)
2. A confounder must be associated with the exposure
3. A confounder must not be an effect of the exposure - or rather it cannot be a causal intermediate

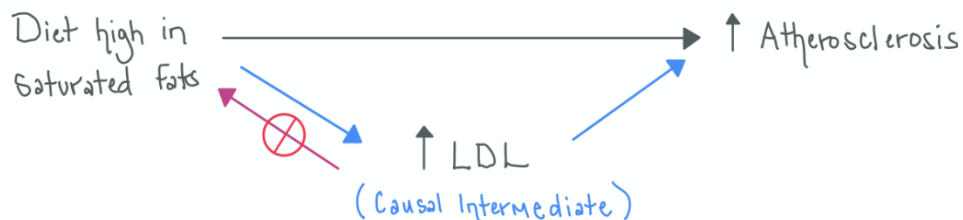


Let's take an example of blood pressure as a risk factor for atherosclerosis. It is commonly known that a chronic elevation in systolic blood pressure can lead to a progression of atherosclerosis. What is another risk factor for atherosclerosis? Well, age is associated with atherosclerosis. The longer we live, the higher our risk is of depositing plaques on our arterial walls. Age is also associated with elevated blood pressure. The older someone is, the more at risk for high blood pressure they are. Therefore, age meets confounding criteria (1) for being associated with the exposure (high blood pressure) and criteria (2) for being associated with the outcome (atherosclerosis). Now for criteria (3). Is age causally associated with elevated blood pressure? No, an elevated blood pressure cannot cause a person to age. Therefore, age also meets criteria (3) for being a confounder – it is not an effect of the exposure. Age is a confounder of the effect of elevated blood pressure on the development of atherosclerosis.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine



What is a causal intermediate? A causal intermediate is a factor that is an effect of the exposure and is an intermediate step in the causal pathway from exposure to disease. Thus, it is associated with both the exposure and the outcome but cannot cause the outcome independent of the exposure. Let's take another example with atherosclerosis as the outcome. A diet high in saturated fats (exposure) can increase a person's risk of developing atherosclerosis (outcome). An elevation in low-density-lipoprotein is also associated with the exposure and outcome in this scenario, but in which direction? LDL is a direct effect of the exposure of eating a diet high in saturated fats and elevated LDL leads to an increased risk of atherosclerosis. Therefore, it is a causal intermediate. Elevated LDL in this example cannot be a confounder despite it being associated with the exposure and the outcome because of the direction of the causal relationship. Having an elevated LDL does NOT cause a person to eat a diet high in saturated fats. LDL does not meet criteria 3 for being a confounder, as it *IS* a direct cause of the exposure.



How does confounding happen? It happens when a third variable, a confounder, that is associated with the exposure and the outcome is unequal between comparison groups in a study. A confounder must have an effect and must be imbalanced between the exposure groups being compared.

For example, if researchers are conducting a cohort study to assess if there is an association between high blood pressure and atherosclerosis, and their exposure group is made up of older individuals and the control group is made up of younger individuals, the observed association between high blood pressure and atherosclerosis may be exaggerated because age is also associated with high blood pressure and atherosclerosis and may distort the observed measure of association.

On the other hand, if patients with high blood pressure were all younger and the control patients were older, you may not observe a true association that exists between high blood pressure and atherosclerosis in the study because the older patients have a higher risk of atherosclerosis than the younger and the younger may not have had enough time to develop as much atherosclerosis as the older population.

Based on these examples, you can see that controlling for confounders in studies is very important. When comparing an exposed group to a control group, they must be equal on confounders in order to observe the true causal association between the exposure and the outcome you wish to study.

Controlling for Confounding in Study Design

There are three ways to control for confounders in study design: (1) Randomization, (2) Restriction, and (3) Matching.

(1) Randomization

Randomization is the act of assigning individuals to exposure or control groups through a random process. This is the benefit of randomized controlled trials. If the study population (n) is large enough, statistically, confounders will be distributed equally between the two groups. While there are other ways to control for confounders as described below, the benefit of RCTs is that confounders we may not be aware of or did not collect data on within the study will be equal between the two groups. This is why the Table 1 of RCTs is so important. If noticeable differences exist between the two arms of the study on something measured, you can control for that factor in the analysis. However, differences between the two groups may draw concern that other differences may exist between the two groups for confounders we are unaware of (for example, something we do not know is a confounder yet because we have not studied it, or something we are unable to take into account because it is unethical or not possible to measure).

(2) Restriction

Restriction is related to eligibility and exclusion criteria in RCTs and cohort studies. It is the limiting of admissibility criteria of subjects into a study, or in other words, confining entry into the study to a group of individuals who fall within a specified category of the confounder. If age is a confounder of concern, investigators may limit the study group to a specific age group, such as age 25-35 in order to study the effect of the exposure on the outcome of interest without risking older age groups confounding the results. If sex is a confounder, you can completely eliminate the variable as a confounder by only admitting females into the study, or vice versa. The goal of restriction is to reduce or eliminate the variability of a confounder between the two groups. A drawback of restriction is it may limit the number of study subjects available to enroll in the study. Similarly, the results of the study may not be generalizable to individuals outside of the demographics of study cohort.

(3) Matching

Matching involves distributing a confounding variable equally between the two study groups. Investigators select study subjects so that potential confounders are distributed in an identical manner between the two groups. For example, in a cohort study assessing the risk of intravenous drug use (IVDU) on acquiring Hepatitis C virus (HCV), a confounder may be HIV. HIV acquisition is associated with IVDU and HCV acquisition, but HIV does not cause HCV. To understand the true measure of association between IVDU and HCV and “controlling” for HIV as a confounder, we could match cases to controls based on their HIV status. To do this, if a 55-year-old male IVDU with HIV and HCV is enrolled in the study as a case, a matched control who is a 55-year-old male IVDU with HIV and no HCV may be enrolled as a control. If a 45-year-old female IVDU without HIV but with HCV is enrolled in the study as a case, a matched control who is a 45-year-old female without HIV and without HCV may be enrolled as a control. Therefore, there will be an equal distribution of individuals in the two groups based on a confounding factor. The drawbacks to matching are investigators cannot use the data to study the association of the matched factor and the outcome. Matching can also be time and resource intensive.

Controlling for Confounding in Analysis

If a research study does not employ randomization, restriction, or matching, it is still possible to control for confounding in the analysis. Two ways to control for confounders in the study analysis is (1) Stratification, and (2) Multivariable methods.

(1) Stratification

After study data has been collected, if data were collected on a potential confounder, it is possible to control for that confounder by stratifying the results. This involves stratifying the study population into subgroups by a confounding variable. This allows the researcher to assess the association between the exposure and outcome in homogenous categories of the confounder. Each stratum should then be free of confounding by the specific stratified variable, essentially making each strata a “restricted analysis.”

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Below is an example from, *Essentials of Epidemiology in Public Health* (Aschengrau and Seage, 2020) from a hypothetical case-control study assessing the risk of DDE exposure (dichlorodiphenyldichloroethylene, the metabolic by-product of the pesticide dichlorodiphenyltrichloroethane [DDT]) on developing breast cancer. A possible confounder in this study may be age. Older patients are more likely to be exposed to DDE because of its increased use and more time for exposure, and older patients are also more likely to develop breast cancer than younger patients. Age is also not on the causal pathway between DDE exposure and developing breast cancer. If age was not equally distributed between the two exposure groups, it may distort the true association.

TABLE 11-5 Crude Data from a Hypothetical Case–Control Study of DDE Exposure and Breast Cancer

DDE level	Cases	Controls
High	500	600
Low	1,500	3,400
Total	2,000	4,000
Odds ratio = 1.9		

If age is a variable for which data was collected during the study, we could stratify the results by age group to see if the observed association remains. When this data was stratified by age younger than 50 years and age older than 50 years, the observed 1.9 increased odds of developing breast cancer due to DDE exposure disappears. Because the OR is 1 for both sub-groups, age was confounding the results and by removing this confounder, we see there is not a true association between DDE exposure and breast cancer.

TABLE 11-6 Age-Stratified Data from a Hypothetical Case–Control Study of DDE Exposure and Breast Cancer

DDE level	Age younger than 50 years		Age 50 years and older	
	Cases	Controls	Cases	Controls
High	50	300	450	300
Low	450	2,700	1,050	700
Total	500	3,000	1,500	1,000
Stratum-specific odds ratio = 1.0			Stratum-specific odds ratio = 1.0	

One drawback to relying on stratification for removing confounding is if the study sample size or power is not strong enough to observe an association after stratification. For example, if this hypothetical study had a smaller proportion of females younger than 50 than over 50 years, the sample size may not be sufficient to avoid random error in the measure of association in the stratum for age younger than 50 years.

(2) Multivariable Methods

When researchers suspect there is more than one confounder affecting the association observed in the study, a simple stratification by one variable will not suffice. A multivariable analysis allows the investigator to control for many confounding variables at once. Essentially, it is a constructed mathematical model that describes the relationship between the exposure, the outcome, and the confounder. Many models exist, and your friendly biostatistician can help you determine which one is best, but essentially the choice of model depends on the relationships between these three variables. A multiple linear regression model is used when the dependent variable is continuous. A logistic regression is used when the outcome is dichotomous. Cox proportional hazard and Poisson models are used when rates from a cohort or RCT are being compared. The drawback of these methods is it requires certain assumptions be made. If these assumptions are not true, the results will be incorrect.

References:

1. Aschengrau, A., & Seage, G. R. (2020). Essentials of epidemiology in public health. Jones & Bartlett Publishers.
2. Rothman, K. J. (2012). Epidemiology: an introduction. Oxford university press.

Submitted 12/2020

III.18 Confounding and Effect Modification- why you need to know (Fili Bogdanic)

If a goal of clinical research is to find real relationships between exposures and outcomes, then an appreciation and understanding of the concepts of **confounding** and **effect modification** are vital. Both effects can lead to erroneous interpretations of data if not properly understood. The two terms are often confused despite being crucially different in that the former must be eliminated for an exposure-outcome relationship to be accurately described, while the latter must be uncovered for that relationship to be accurately described. In the following chapter, you'll learn the difference between the two and how stratified analysis can help reveal their presence.

First, a couple definitions to note:

- A **measure of association** is simply the calculated effect of an exposure on an outcome. This is commonly expressed as a relative risk (RR).
- The **crude** measure of association is the initial association calculated before subgroup analysis. If either confounding or effect modification are present, this crude measure of association will be misleading.
- The **adjusted** measure of association is what is calculated after stratified analysis (or subgroup analysis) is done. In terms of RR, comparing the adjusted RR with the crude RR can help reveal whether confounding or effect modification are present.

Confounding, as discussed in a prior chapter, is a distortion of a measure of association between an exposure and an outcome. If the goal is to conduct high quality research, then it is important to correct for and eliminate confounding as much as possible in order to obtain clarity on an exposure-outcome relationship.

Key Points about Confounding:

- The confounding factor is associated independently with both the exposure and the outcome.
- Its association with the exposure and the outcome is **not** an intermediate step of a causal pathway between them.
- Stratification of the results into subgroups by the confounding factor exposes the effect of the confounding factor.

Example of Confounding:

As a quick example, think of a study that looks at the association between motorcycle riding and lung cancer. In this hypothetical study, an increased crude measure of association of lung cancer was found in people who regularly ride motorcycles as compared to those that do not ride them.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Exposure (motorcycle riding)	+ Outcome (cancer)	- Outcome (no cancer)
Riders	100	900
Non-Riders	25	975

Crude RR: 4.00 (95% CI, 2.60 to 6.15)

However, in this hypothetical population, people that rode motorcycles were also significantly more likely to smoke than non-motorcycle riders. When results were stratified into the subgroups of “smokers” and “non-smokers”, there were no differences in lung cancer rates between people who ride motorcycles and people who don’t.

Exposure (motorcycle riding)	+ Outcome (cancer)	- Outcome (no cancer)
Smoking Riders	97	403
Smoking Non-Riders	23	77

Smoking adjusted RR: 0.84 (95% CI, 0.56 to 1.26), thus no significant difference

Exposure (motorcycle riding)	+ Outcome (cancer)	- Outcome (no cancer)
Non-Smoking Riders	3	497
Non-Smoking Non-Riders	2	898

Non-Smoking adjusted RR: 2.70 (95% CI, 0.45 to 12.01), again, no significant difference

The variable of smoking was the confounding factor in this study because it caused a distortion of the apparent association between motorcycle riding and lung cancer rates; we can see this because the initial crude RR appeared statistically significant, but neither of the adjusted RRs were significant.

It is worth pausing to consider the relationships between the exposure, the outcome, and the confounding factor. As mentioned previously, a confounding factor is independently associated with both the exposure and the outcome but not as a causal intermediate. In this example, smoking was independently linked to riding motorcycles (behaviorally, people who engaged in one type of risky behavior were also more likely to engage in other types) as well as developing lung cancer (via directly damaging the lungs). Motorcycle riding and lung cancer had no real association once the confounding factor was removed with stratified analysis. (A >10% change from the crude RR to the adjusted post-stratification RR in the subgroups is generally accepted as the statistical benchmark for defining a confounding factor.)

Effect modification is different than confounding in several important ways. On the most basic level, effect modification describes a situation where an association between an exposure and an outcome

differs depending on a third variable called the modifier. Unlike confounding which obscures true associations (or lack-thereof), effect modification is of scientific interest, and identifying the modifiers of an exposure-outcome relationship helps clarify true associations and mechanisms.

Key Points about Effect Modification:

- The modifier is associated with the outcome but not the exposure.
- It “modifies” the causal link between the exposure and the outcome (depending on the degree or quality of the exposure).
- The presence of effect modification can be uncovered by stratified analysis.
-

Example of Effect Modification:

Imagine a study looking at aspirin and Reye’s syndrome, characterized by post-viral brain and liver dysfunction. The study looks at a large population of patients under 30 years old and concludes, based on the crude measure of association, that there is no association between exposure to aspirin during a viral illness and Reye’s syndrome. (Note: in this example, the numbers do not reflect real-life rates of Reye’s syndrome which is rare!)

	+ Outcome (Reye’s)	- Outcome (no Reye’s)
Aspirin use during viral illness	15	9,985
No aspirin use	6	9,990

Crude RR: 2.50 (95% CI, 0.97 to 6.43), hence the result is non-significant despite the large #s

However, when the data are stratified by age, the adjusted RR in the subgroup for ages 0 – 10 years old reveals a significantly increased RR of Reye’s syndrome. In the subgroups for other age groups, no such increased adjusted RR is seen. Therefore, it is concluded that young age is a modifier in the relationship between aspirin exposure and the development of Reye’s syndrome.

AGES 0 – 10:

	+ Outcome (Reye’s)	- Outcome (no Reye’s)
Aspirin use during viral illness	14	2986
No aspirin use	2	2998

Adjusted RR: 7.00 (95% CI, 1.59 to 30.77)

AGES 10 – 20:

	+ Outcome (Reye’s)	- Outcome (no Reye’s)
Aspirin use during viral illness	1	2999
No aspirin use	2	2998

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Adjusted RR: 0.50 (95% CI, 0.05 to 5.51)

AGES 20 – 30:

	+ Outcome (Reye's)	- Outcome (no Reye's)
Aspirin use during viral illness	0	4000
No aspirin use	2	3998

Adjusted RR: 0.20 (95% CI, 0.0096 to 4.1648)

As you can see from the data, only the adjusted measure of association for the “ages 0 – 10” group is significant; the others are not.

To better understand how young age is a modifier here and not a confounder, think back to the prior example of confounding. In that case, the confounding factor—smoking—was associated both with motorcycle riding (via predisposition for risky behaviors) and lung cancer (via direct damage to lungs). When the results were stratified by smoking status however, the adjusted RRs were similar in the two groups. In this example of effect modification however, when the groups are stratified into subgroups based on the modifier—in this case, age groups—then a significant difference in adjusted RRs is revealed. In this way, age is said to *modify* the risk for Reye's syndrome, making it more likely if a person is a child and less likely if they are an adult. It is important to note that the modifier—young age—is only associated with the outcome of Reye's syndrome; it is not associated with aspirin use in any meaningful way.

Designing Studies to Address Confounding and Effect Modification:

Stratified analysis as a means of assessing for confounding or effect modification is particularly relevant in clinical medical research because pathophysiology is rarely ever as simple as a single exposure influencing a single outcome. Considering this, it is important that before a study is undertaken, a very thorough background review of the existing literature is done so that subgroups can be identified and prespecified, and to ensure the right data is collected at the onset. This is easier said than done, as it is not always obvious beforehand which covariables are going to be relevant. You can imagine that very new fields or topics of study are particularly vulnerable to omitting relevant subgroups analyses—and therefore are more susceptible to confounding or failing to reveal effect modification where it is present—if a dearth of data exists to inform these decisions.

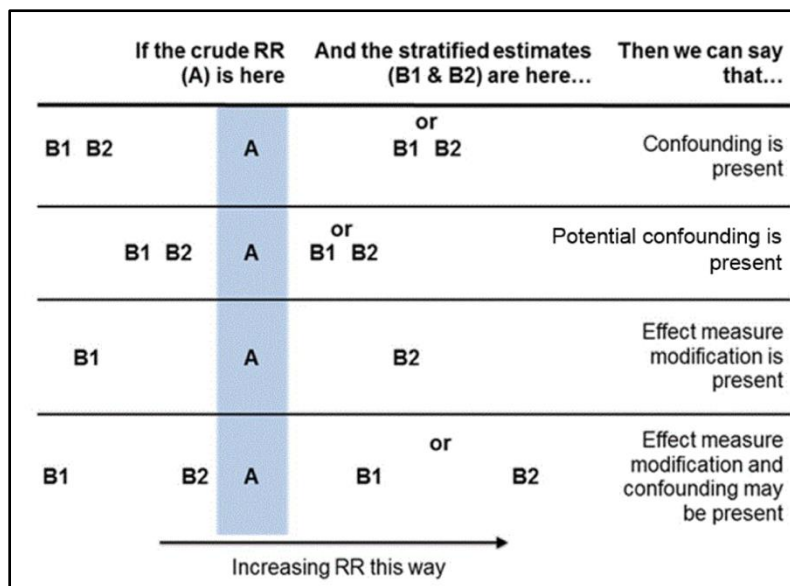
When reasonable subgroups are chosen however, the expected number of participants in each subgroup must be considered because of how it will impact statistical power. You may already be aware that as the N of a study increases (often in EBM, N is the number of patients or patient encounters), then the study's power also increases. As a reminder, the power (expressed as $1 - \beta$) is the probability that the study will correctly reject the null hypothesis when a specific alternative

hypothesis is true. In other words, it expresses the probability of truly detecting a real association between exposure and outcome if one exists. During subgroup analysis, the N in each subgroup will obviously be less than the total N. Practically, this means that interpretations of whether a result is significant or not may be obscured if a result is not adequately powered. Often, published articles will include a disclaimer somewhere in the results section stating that the subgroup analyses were not adequately powered to reveal true significant results but are included for interest.

In Summary: How stratified analysis differentiates confounding and effect modification:

- If **confounding** is present, the degree of association between the exposure and the outcome will be similar between the stratified groups, but will differ from the crude measure of association (of the initial, pooled results).
- If **effect modification** is present, subgroup analysis will reveal a significant measure of association, when the results are stratified by the modifier (in the above example, age). In simple terms, a previously hidden effect will be revealed after stratification.

The following diagram is a helpful visual of these trends:



(From the Second Edition ERIC Notebook, UNC Department of Epidemiology)

In this visual, the crude RR is shown as A, while the adjusted RRs in two subgroups are shown as B1 and B2. Note that in the rows where confounding is depicted, B1 and B2 are similar to each other and fall on the same side of A. This is consistent with what we know about confounding, that when the confounder is corrected for (with stratification), the subgroups will show a similar relationship between exposure and outcome (either more or less than what was seen with the confounder influencing the crude RR, in this case, A). On the other hand, in the rows where modification is present, B1 and B2 are separated on either side of A. This is consistent with what we learned about

modification, namely that it reveals a difference in subgroups where it was previously hidden (before stratification).

Notes: (a) In the EBM Database, outcomes are reported as relative risk reduction (RRR) or relative risk increase (RRI), not relative risk or risk reduction (RR). Crude rates are calculated by the reviewer, and are presumed to have had confounding reduced or eliminated as the vast majority of studies are high quality randomized RCTs. When the unadjusted and the adjusted risks do not agree, confounding may be present.

(b) Since randomization requires that the exposure status of individuals be assigned to study participants, observational study designs such as cross-sectional, cohort, case-control and ecological studies cannot use randomization to control for confounding. For controlled clinical trials however, randomization is a common method which attempts to control for confounding.

References:

- Alexander LK et al. "Confounding Bias, Part II and Effect Measure Modification." ERIC Notebook, 2nd edition. UNC Gillings School of Global Public Health Online.
- LaMorte WW & Sullivan L. "Confounding and Effect Measure Modification." Boston University School of Public Health Online.

Submitted 12/2022

III.19 Coming to Terms with Composite Measures (Mallory Perez, GSM4)

Is the whole really greater than the sum of its parts?

Description of Composite Measures

Clinicians, institutions, and policymakers measure and report outcomes in clinical trials to inform medical decision-making for individual patients or populations.¹ The selection of a valid primary endpoint is critical for RCTs to demonstrate the efficacy of their interventions. Composite measures are widespread and are more frequently becoming the primary endpoint in RCTs, especially prominent in cardiovascular literature. A composite measure (also referred to as "composite endpoint") is a combination of a number of individual measures into a single measure that results in a single score. There are strengths and drawbacks to this approach.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Let’s walk through an example from the FOURIER trial (2017).² Suppose an investigator wants to know the effect of evolocumab, a monoclonal antibody that inhibits PCSK9 and lowers LDL, on cardiovascular events. With the intent to use this drug for prevention, the investigator specifically examines how much overall benefit a patient may receive from this drug. Thus, the primary endpoint is defined as a composite of cardiovascular death, MI, stroke, hospitalization for unstable angina, and coronary revascularization, rather than as individual endpoints. The null hypothesis is the composite outcome for patients receiving evolocumab will be no different than that for patients on placebo. See Table I below for additional examples.

Table I. Examples of composite measures

Field	Trial	Composite	Components
Cardiology	SYNTAX (2009): percutaneous coronary intervention v CABG for severe CAD ³	Major adverse cardiac or cerebrovascular event	<ul style="list-style-type: none"> • Death from any cause, • Stroke, • Myocardial infarction, or • Repeat revascularization
Neonatology	CAP (2006): caffeine v placebo ⁴	Short-term outcomes for newborns with apnea of prematurity	<ul style="list-style-type: none"> • Death, • Cerebral palsy, • Cognitive delay, • Deafness, or • Blindness
Transplant	Fan et al. J Hematol Oncol. (2017): G-CSF-mobilized peripheral blood cells v. G-CSF-primed bone marrow transplants in adult leukemia patients ⁵	GVHD-free/relapse-free survival (GRFS)	<ul style="list-style-type: none"> • Absence of the following: • Grade 3-4 acute GVHD, • Systemic therapy-requiring chronic GVHD, • Relapse, or • Death
OB/GYN	PREMODA (2006): planned cesarean v vaginal delivery of term breech births ⁶	Morbidity / mortality	<ul style="list-style-type: none"> • Fetal mortality • Neonatal mortality • Severe neonatal morbidity

There are many types of composite measures. For brevity, we will mention three: indexes, scales, and typologies.⁷

- Indexes create an aggregate score from individual attributes of various variables.⁷The examples in Table I are indexes.
- Scales (e.g., Likert scale) analyze any logical or empirical intensity structures that exist among a variable’s indicators in a graded fashion.⁷

- Typologies are nominal composite measures often used in social research, most effective when interpreted as an independent variable.⁷

Let's focus on indexes as these are most often utilized in RCTs. The key steps in development of a composite measure are: selection of candidate measures (i.e., individual outcomes, component variables) for inclusion in the composite, examination of the empirical relationships between components, scoring the components, and validating the methodology for composite measure calculation.⁷

Development of Composite Measures

In order to determine which type of composite measure is best suited for your study, you have to select your component variables and know the relationships between them. Technical and clinical experts usually form teams to review and select component variables by means of consensus, imbuing the composite measure with face validity. The goal of the team is to select components that ensure the composite upholds the primary objective of the trial, is biologically plausible, and represents a construct that is meaningful to clinicians and patients. A well-developed RCT will pre-specify the criteria for inclusion of individual measures and the study's process for measure selection.

Common criteria for component variable selection are as follows:

- face validity,
- unidimensional,
- degree of specificity attainable for measuring the desired dimension, and
- amount of variance provided by the component.

If individual measures represent unique aspects of the composite, then these components should be related empirically. Examining these relationships is an internal validity check and must be carried out before moving to the next step to limit duplications and contradictions.

Once the included measures are determined and their empirical (bivariate and multivariate) relationships have been examined, the researchers establish the ordinal hierarchy among the components, if such relationships exist.

Now, it's time to decide how the component measures will be "rolled up" into one composite score. This step involves developing a weighting scheme and a scoring strategy.

Composite measure calculation methods may include:

- all-or- none/any-or-none scoring at the patient level (see Table I examples above),
- sum,
- average,
- weighted average, or
- opportunity scoring.

The success of this step relies heavily on input from patients and clinicians regarding what is most meaningful to them. The scoring strategy may consider the range of scores investigators hope to achieve across measured entities (e.g., patients, clinicians, hospitals).

Evaluating the statistical significance between the composite score in the intervention and control groups can be a challenge because the p value set for the individual measures may not carry over in a 1:1 manner. This is especially important because it influences the number of patients needed to ensure the study is adequately powered. Suppose we apply an all-or-none scoring approach to individual measures determined to be of equal importance/weight. Several hypotheses (one per measure) will be simultaneously tested at significance level α ; however, the probability of falsely rejecting at least one of the null hypotheses is, in general, no longer controlled at this level. This phenomenon is known as inflation of the type I error; it would be inappropriate to apply a “global significance level.” Statisticians utilize multiple testing of composite null hypotheses to address this issue (Table II). If one assumes a priori there is a common effect size δ , then assigning equal Type I error probability to each test is reasonable. Conversely, if some effect sizes were believed a priori to be larger than others, one assigns greater Type I error probability to the tests with larger effect sizes.^{8,9}

In short, calculation of p values and desired sample sizes for studies with primary composite endpoints is complex; researchers are responsible for detailing their approach in their methods for the reader’s critical appraisal.

Table II. Multiple testing scenarios to improve the interpretation of composite endpoints⁹

Multiple testing procedure	Clinical study situation	Adjusted local significance levels	Guide – how to use	Visualization
Hierarchical testing	Given is a composite endpoint with equal or unknown treatment effects in its components	No adjustment necessary	Order of the component hypotheses according to the expected effect sizes starting with the largest. Begin by testing the composite endpoint followed by the ordered components. Stop testing if a null hypothesis cannot be rejected.	See Fig. 1
Hierarchical testing with Bonferroni-Holm	Given is a composite endpoint with different treatment effects in its components	α for the composite endpoint; for the components compare smallest p-value to $\frac{\alpha}{n}$, compare next smallest p-value to $\frac{\alpha}{n-1}$, ..., compare largest p-value to α	Begin testing the composite endpoint. If this null hypothesis is rejected, test all component hypotheses (in the order of the p-values starting with the smallest)	See Fig. 1
Intersection-union-test	A successful outcome should be based on a significant effect in the composite endpoint <i>and</i> in the most harmful component	No adjustment necessary	Test the composite endpoint and the main component; only if both null hypotheses are rejected the result is satisfying	See Fig. 2
Union-intersection-test	Two different composite endpoints are under consideration; a successful outcome should be based on a significant effect in at least one of them	For the two endpoints compare the smaller p-value to $\frac{\alpha}{2}$ and compare the larger p-value to α	Test both composite endpoints; if at least one null hypothesis is rejected, the result is satisfying	See Fig. 2

Finally, researchers seek external validation for the composite measure by comparing scores to other indicators of the variable, not included in the measure, if available.

Benefits of Composite Measures

At this point, you may be thinking, “Wow, that seems like a lot of work. Why bother using composite measures at all?” Composite measures in RCTs have numerous benefits.

As we have garnered and applied more evidence for clinical decision-making, outcomes have improved. Conventionally measured outcomes like mortality have become less frequent occurrences for many common conditions and procedures. Consequently, the selection of relevant clinical outcomes to target requires increasingly more creativity. Combining individual measures into a single composite typically increases event rates and improves detection of variance (increased statistical precision). Thus, composite endpoints are one solution for maintaining the feasibility of an RCT, particularly in the setting of low event rates, high cost, and long follow-up. With a composite endpoint, the RCT can include fewer patients, which is more likely to result in an adequately powered study, not to mention the time and cost savings.

Composite measures allow researchers to avoid choosing between several important outcomes, particularly for diseases with multidimensional presentations. In theory, these measures offer the opportunity to summarize multiple dimensions of a concept more comprehensively and succinctly than reporting individual outcomes. Moreover, composite measures are believed to be able to better capture a latent, unmeasurable endpoint better than a single measure.

Composite measures also have the added benefit of being able to customize and validate the methods for measure calculation through the design of an approach that allows for ordinal rankings with a preset, desired range of variation.¹⁰⁻¹³

Risks of Composite Measures

On the flip side, the use of a composite measure is not without risks, the greatest of which is arguably misinterpretation of results by clinicians, policymakers, and patients. Greater precision has to be weighed against greater uncertainty.

The value of the composite measure relies on the validity of its components. The outcomes that contribute to a composite measure must be "*associated with the primary objective.*" The ideal components are similar, not identical. A composite endpoint made up of component measures with large variability in importance to patients or clinicians is concerning. If the component measures are of equal (or relatively similar) importance, the distribution of the relative risk reduction among the components is less concerning because if the composite crossed the threshold for statistical significance, one can be assured that an important component played a substantive role. The larger the gradient in importance between component measures, the greater the concern for using the composite to inform decision-making. Similarly, if when applying a weighting scheme, the more important components occur far less frequently than the less important components, the utility of the composite is once again limited.

The minimum clinically important difference (MCID) for a composite outcome is hard to define. Even if the occurrence and importance of components is known, the overall occurrence and importance of the composite outcome is very difficult to estimate.

By aggregating event rates, some argue measurement of the treatment effect is diluted. Composite endpoints with an all-or-none approach such as time-to-first-event variables only consider the first occurring event, and the rich data that comes from tabulating the number and severity of events is lost in the composite score.

While an intervention may result in a statistically significant difference compared to the control at the level of the primary composite endpoint (e.g., adverse cardiovascular event), an individual outcome (e.g., mortality), presented as a secondary outcome, may show no statistically significant difference. Therefore, naming the composite as a list of the individual outcomes may be confusing & inaccurate when taken out of context.¹⁰⁻¹³

Reader Strategies

When appraising studies with composite endpoints, the reader should inquire whether or not the composite endpoint is an appropriate basis for medical decision-making. The questions in the table below help guide this determination. If “yes,” the reader can take greater confidence in using the treatment effect on the composite endpoint as the basis for medical decision-making. If “no,” individual measures likely provide more informative data.¹

Evaluating the utility of a composite endpoint for medical decision-making	
1.	Are the component endpoints of similar importance to patients?
2.	Did the more or less important endpoints occur with similar frequency?
3.	Can one be confident that the component endpoints share similar relative risk reductions? <ul style="list-style-type: none">• Is the underlying biology of the component endpoints similar enough such that one would expect to see similar relative risk reductions?• Are the point estimates of the relative risk reductions similar, and are the confidence intervals sufficiently narrow?

Figure II. Questions to Aid Clinicians in Evaluating the Utility of a Composite Endpoint ¹

Author Strategies

- Explicitly state the primary objective of the study. Reference the objective throughout development and execution of the trial.
- Ensure the composite endpoint is useful for clinicians’ decision-making.
- Define the composite endpoint to be specific to an overall disease process to minimize the likelihood of misinterpretation or contradictions in the directions of individual measure effects.
- Clearly and carefully specify the individual measure selection process. Only include component measures based on evidence-based, valid, and reliable data.

- Include individual component measures as secondary outcomes.⁹
- Decompose the individual measure effects. Provide results of analyses for individual outcomes (preferably in table format) and explain discrepancies in significance.⁹
- Present and empirically test methods used for weighting and combining individual measures into the composite endpoint.
- Name the composite purposefully in order to limit misinterpretation. Rather than defining a composite as a list of its components (e.g., death, recurrent infarction, and stroke), a more generalized description may be more appropriate (e.g., “major adverse cardiac and cerebrovascular event”).

Conclusion

Upon deciding that a study is well-suited for use of a composite measure, researchers need to identify the components to be used in the construction of the composite measure and develop the methods (e.g., weighting, scoring) to create and validate the composite measure. As RCTs utilize more and more composite endpoints, knowledge of the benefits and risks of these measures becomes increasingly important. When an RCT utilizes a composite endpoint, clear and specific documentation of the trial’s methods and results is critical. Ongoing research is needed to determine how to optimally communicate with readers to mitigate the greatest risk of composite measure use, misinterpretation. This chapter provides strategies for the reader & the author as they engage with trials using composite measures.

References:

1. McCoy CE. Understanding the use of composite endpoints in clinical trials. *West J Emerg Med.* 2018;19:631–4.
2. Sabatine MS, Giugliano RP, Keech AC, Honarpour N, Wiviott SD, Murphy SA, Kuder JF, Wang H, Liu T, Wasserman SM, Sever PS, Pedersen TR. FOURIER steering committee and investigators. Evolocumab and clinical Outcomes in patients with cardiovascular disease. *N Engl J Med.* 2017;376:1713–22.
3. Serruys PW, Morice MC, Kappetein AP, Colombo A, Holmes DR, Mack MJ, Stahle E, Feldman TE, van den Brand M, Bass EJ, et al. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. *N Engl J Med.* 2009; 360:961–97.
4. Schmidt B, Roberts RS, Davis P, Doyle LW, Barrington KJ, Ohlsson A, et al. Caffeine therapy for apnea of prematurity. *N Engl J Med.* 2006;354:2112–21.
5. Fan Q, Liu H, Liang X, Yang T, Fan Z, Huang F, Ling Y, Liao X, Xuan L, Xu N, et al. Superior GVHD-free, relapse-free survival for G-BM to G-PBSC grafts is associated with higher MDSCs content in allografting for patients with acute leukemia. *J Hematol Oncol.* 2017;10:135.
6. Goffinet F, Carayol M, Foidart J-M, Alexander S, Uzan S, Subtil D, et al. Is planned vaginal delivery for breech presentation at term still an option? Results of an observational prospective survey in France and Belgium. *Am J Obstet Gynecol.* 2006;194: 1002–1011. pmid:16580289

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

7. Babbie, E. (2012). Chapter 6: Indexes, Scales, and Typologies. In *The practice of social research* (13th ed., pp. 157-175). Cengage Learning.
8. Song MK, Lin FC, Ward SE, Fine JP. Composite variables: when and how. *Nurs Res.* 2013;62(1):45–9. doi: 10.1097/NNR.0b013e3182741948.
9. Schüler S, Mucha A, Doherty P, Kieser M, Rauch G (2014) Easily applicable multiple testing procedures to improve the interpretation of clinical trials with composite endpoints. *Int J Cardiol* 15:126 – 32.
10. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *JAMA.* 2003;289(19):2554–2559. doi:10.1001/jama.289.19.2554
11. Ross S. Composite outcomes in randomized clinical trials: arguments for and against. *Am J Obstet Gynecol.* 2007;196(2):199–16.
12. Rauch G, Rauch B, Schüler S, Kieser M. Opportunities and challenges of clinical trials in cardiology using composite primary endpoints. *World J Cardiol.* 2015;7:1.
13. Barclay M, Dixon-Woods M, Lyratzopoulos G. The problem with composite indicators. *BMJ Quality & Safety* 2019;28:338-344.

Submitted April 22, 2020

III.20 Evaluation of Screening Tests (Alex Fiorentino)

Types of Tests

As clinicians, we order tests with a variety of different purposes in mind. For example...

Diagnostic testing occurs when an individual patient presents with a problem – a symptom, physical exam sign, or abnormality on prior workup. The goal of a diagnostic test is to clarify the cause or nature of the patient’s problem (including ruling in or ruling out individual etiologies) in order to improve the quality or quantity of life for the *individual patient* being tested.

In contrast, *Screening* occurs when entire populations of asymptomatic patients are tested for pre-clinical disease or disease risk factors. The goal of a screening test is to improve quality or quantity of life for the entire *population* being screened. That said, individual patients are screened for pre-clinical disease or disease risk factors.

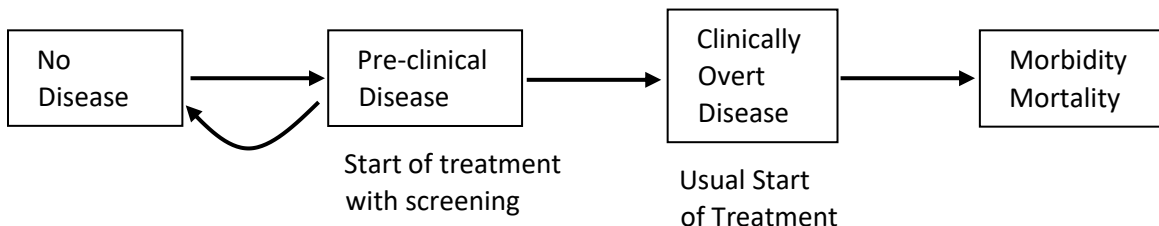
Screening and Prevention

A screening test may serve as *primary prevention* when the test identifies risk factors for disease (e.g., hyperlipidemia) and directs treatment to prevent disease from developing (e.g., cardiovascular disease). A screening test may also serve as *secondary prevention* when the test identifies pre-clinical disease (e.g., mammographic identification of breast cancer) and allows early treatment. Some screening tests span both primary and secondary prevention (e.g., Pap testing for early detection of cervical neoplasia with HPV co-testing for detection of high-risk HPV infection).

Screening and Natural History

When we screen for a disease, we implicitly make several assumptions about the natural history of the disease (see figure below).

1. There exists a pre-clinical phase when the disease is not clinically apparent but can be detected with testing.
2. Without intervention, the pre-clinical disease proceeds to overt disease and morbidity/mortality frequently enough to warrant intervening on everyone who develops pre-clinical disease.
3. Early treatment starting in the pre-clinical phase is effective for reducing the development of morbidity and mortality and this benefit outweighs the harms of early diagnosis and treatment.



If any of these conditions are not met, screening is unlikely to be successful. For example, in adolescent women, 90 to 95% of low-grade cervical lesions detected on Pap testing will regress spontaneously without intervention, so the benefits of cervical cancer screening in this age group do not outweigh the harms.

Challenges in Evaluation of Screening Programs

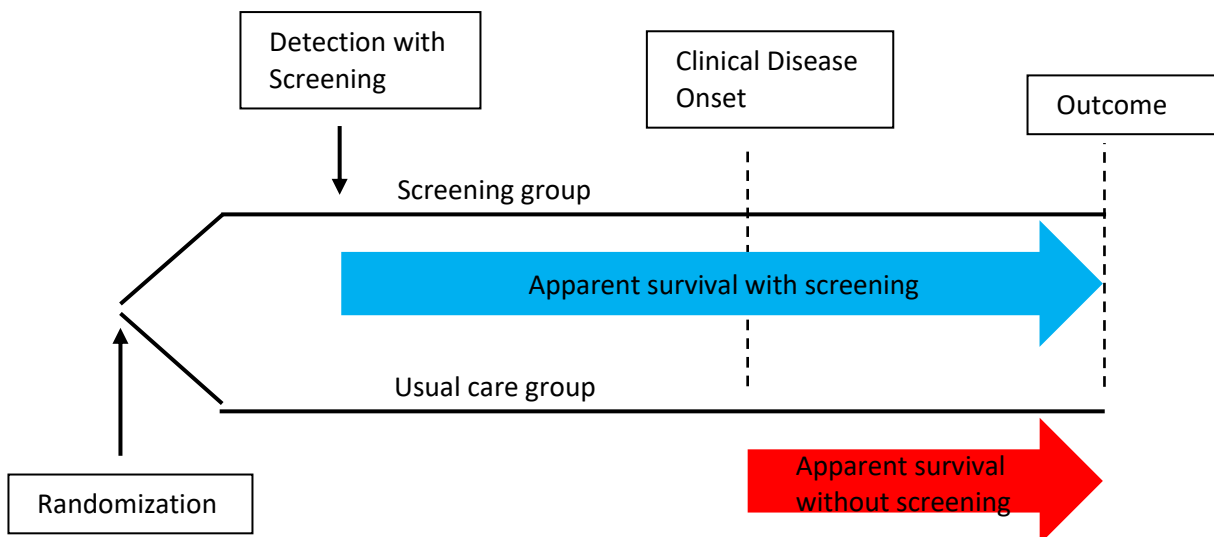
Like other interventions, screening programs can be evaluated with a variety of study designs (cohort studies, randomized controlled trials, etc.). Some of the challenges in interpreting these studies are reviewed below.

1. Volunteer (selection) bias

One way to evaluate a screening program is with a cohort study, in which screening is offered to a general population and screening and non-screening cohorts are monitored for mortality, quality of life, or other outcomes. However, these two cohorts are inherently dissimilar with respect to whether they volunteered for the screening program (selection bias), and this difference may be a powerful prognostic factor. For example, in the Health Insurance Program (HIP) study, an early RCT examining the effectiveness of mammography, some women were randomized to mammography and some to no mammography. In follow-up, mortality in women who were randomized to mammography but refused to undergo it was 96 per 10,000 person years, whereas mortality in women who were randomized to no mammography was much lower at 75 per 10,000 person years, even though both of these groups ultimately had the same exposure (none) to the screening intervention. As with other types of clinical evidence, randomization is the gold standard for screening program evaluation.

2. Lead time bias

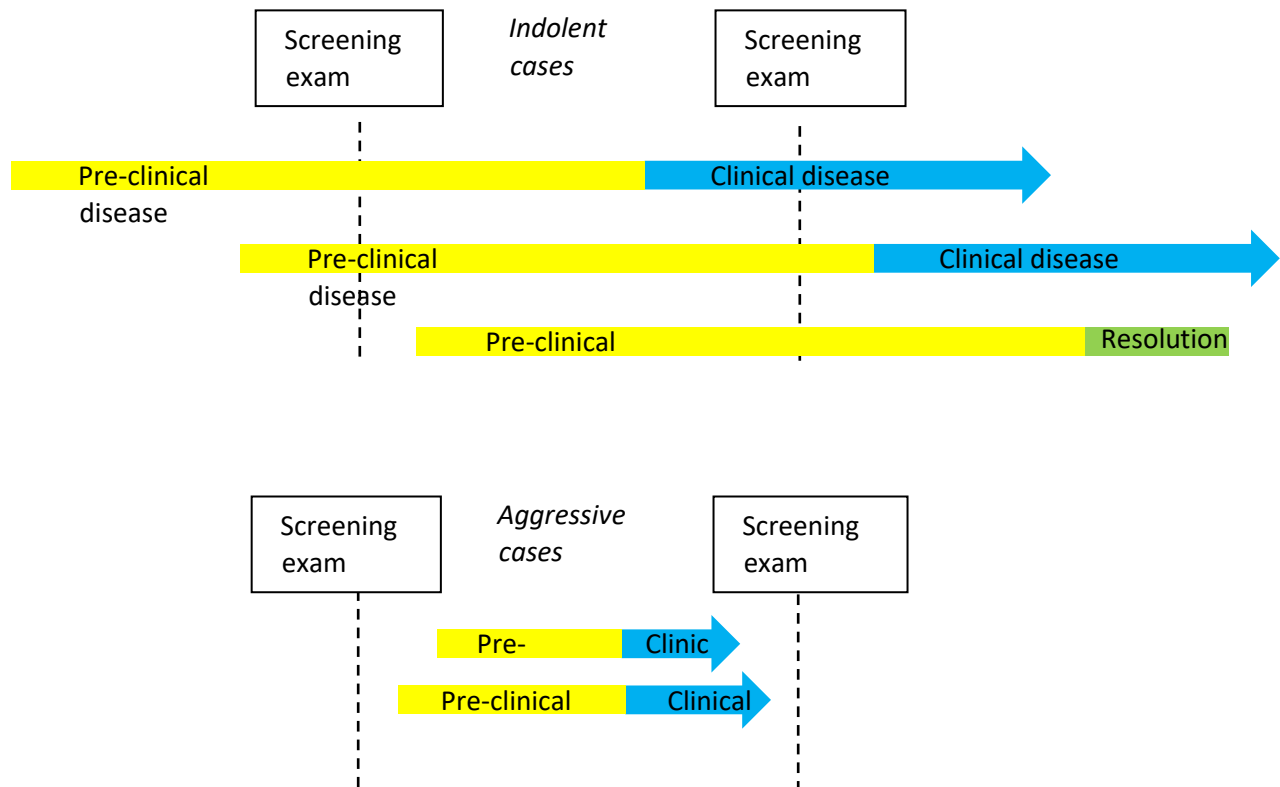
Even in randomized controlled trials, the way in which we measure the effectiveness of a screening program can produce the illusion of benefit where there is none. Imagine that we have introduced a new cancer screening program and want to test its effectiveness. We compare 5-year survival in two groups: one randomized to screening and one randomized to usual care.

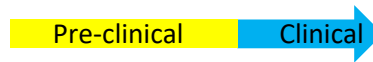


Because it detects disease earlier, the screening test will inherently increase apparent 5-year survival from the time of diagnosis even if the patients who undergo screening do not live any longer than those who are not screened. This is called *lead time bias*. One strategy for dealing with lead time bias is to use outcomes that are not susceptible to this bias. For example, if we compared disease-specific mortality in the *entire screened and unscreened populations*, rather than 5-year survival or case fatality rate, then the study would more directly assess the success of the screening program at a population level and would not be susceptible to lead time bias. Another strategy would be to estimate the lead time produced by the screening intervention on average and subtract this from the 5-year survival in the screened group; however, this strategy would then be susceptible to bias and error in lead time estimation.

3. Length bias

What if a screening program selectively identifies mild cases of disease and misses more aggressive cases? In fact, this will nearly always be the case. To see why, imagine we are evaluating a new cancer screening program at our institution. We randomize one group to screening every 5 years and the other to no screening (see figure). In more indolent cases where the pre-clinical phase lasts >5 years, our test will succeed in pre-clinical detection 100% of the time. In more aggressive cases where the pre-clinical phase lasts only 2 years, we will succeed in pre-clinical detection only a fraction of the time, with the remainder of these aggressive cases becoming clinically overt in between screening exams.





Because rapidly progressive disease is harder to detect with periodic screening, the cases that our screening program identifies will be, on average, more indolent than the average case in the entire population. This is called *length bias*, and it represents a form of selection bias. These indolent cases will likely have reduced morbidity and mortality. If we assess our screening program by comparing the case fatality of those who screened positive in our screening group to the case fatality of those who were diagnosed with clinical disease in our non-screening group, it will appear that our screening test improves survival, even if it has no effect whatsoever. As with lead time bias, the best way to manage length bias is to avoid using case fatality or 5-year survival as primary outcomes and instead compare overall or disease-specific mortality in the *entire screened and unscreened study populations*.

Submitted May 2019

III.21 Cancer Screening: Elements of an effective screening tool and other considerations- (Eric Jayne, GSM4)

i. Introduction

Rebecca (name changed) found herself in a situation after she underwent routine colonoscopy at the age of 75 years. Aside from well-controlled hypertension, she was quite healthy, was still playing tennis every day and enjoyed an active lifestyle. As an unfortunate and rare complication, she suffered a colonic perforation during this procedure, which ultimately led to surgery and placement of a colostomy. During recovery she became depressed and for months was unable to do the things in life that she previously enjoyed. Sadly, just over a year after her colonoscopy, she passed away from a peritoneal bacterial infection.

While the anecdote above is unique to Rebecca and her life, it encapsulates an all-too-familiar situation that we encounter in healthcare – when routine care or procedures result in unexpected or horrific outcomes. At the advent of our training, we learn that our interventions always come with risks, and while we strive to mitigate these risks, they can be devastating to patients and their families when they occur. The purpose of this chapter is not to argue against cancer screening, but rather to remind ourselves of the risks that are associated with screening, thereby strengthening our framework for assessing its value and when making shared decisions with the individual patient.

ii. **What Makes a Good Screening Test?**

Screening for cancer involves assessing the benefits, harms, and impact of the screening intervention. What are the important characteristics of an effective screening test for the general population? As covered in section III.20, the goal of a screening test is to improve the health status of the entire screened population. While we normally conceptualize the value of screening tests based on their utility in detecting disease, in practice, these tools derive value in the context of a broad range of measures: accessibility, cost, rates/severity of harms, and clinical impact on disease course (just to name a few):

Accessibility – Without equitable access, a screening tool may be unavailable to large or particular portions of a population, thereby decreasing its efficacy for improving the health of the entire population. Not only is this problematic from a health justice perspective, but it also may render the intervention ineffective if we are to be inclusive.

Cost and Simplicity – Without taking a deep dive into the complex topic of costs in healthcare it should be evident that if a screening test does not provide cost-effective outcomes (compared to alternative strategies), it will not be a feasible intervention for the system to support. Likewise, if it is too complicated or burdensome to incorporate into clinical practice, it will not be adopted by patients and clinicians.

Clinical Impact – Remember, screening tests are for *asymptomatic* patients, i.e. they have to detect disease before it would be clinically apparent and thereby create the opportunity to alter the disease course ahead of when it would otherwise manifest itself in the form of symptoms or signs. The ability of a screening test to alter the clinical course also relies on the existence of an effective intervention. It's no good starting the treatment for something earlier if there is no possibility of an improved outcome. And while one may argue that there is still value for knowing about something even if we can't do anything about it, this is typically not cost-acceptable or indeed equitable from the perspective of the system.

Benefits that outweigh the harms – This idea is of course a sub-category of clinical impact described in the last paragraph, but emphasizes that if a screening test has true benefit for some patients, but an equal or greater amount of harm for another group of patients, it may not be an acceptable intervention from a population perspective.

High Sensitivity (and hopefully high specificity, too!) – As screening tests are used to interrogate for diseases of which prevalence is typically very low, a good screening test must have a high sensitivity so that it does not miss the few cases of disease present (particularly, *early* in disease course – as above).¹ Of course, high sensitivity often comes with the price of lowered specificity and increased rates of false positives. Similarly, good screening tests tend to have high negative predictive value (NPV) and low positive predictive value (PPV). Thus, if a test is negative, you want to be confident that

it is a true negative (high NPV), whereas if a screening test results positive, it can be confirmed with tests that have a higher specificity. See section I.2 (Introductory Essentials) for a review of these topics.

Bias in interpreting survival analyses- recall that bias in interpreting the impact of tests may be at play. For example, lead-time bias can occur when a disease is detected at an earlier time point than it would have been if it had been diagnosed by its clinical appearance. This can lead to an overestimation of survival duration. Alternatively, length-time bias can lead to an overestimation of survival due to the relative excess of cases detected that are slowly progressing. Both of these cause overestimation of survival duration, and hence of efficacy of the screening test. An example of the former (lead-time bias) is the impact on mortality (it falls) from colon cancer when the recommended age to start screening moves from 50 to 45y. An example of the latter (length-time bias) can occur when prostate cancer patients who are seen over a month in clinic may be more likely to progress slowly than those with aggressive prostate cancer.

Consider one last point before moving onto the next section in which we'll take a look at the specifics of screening tools we actually use. It's important to remember that the vast majority of patients who undergo population cancer screenings see *neither* benefit *nor* harm; by far the most common outcome for individual patients is a normal screen. This is one reason why RCTs assessing screening tools are hard to design; they require huge sample sizes to acquire the power necessary to detect significant differences. One example of this is a trial that evaluated prostate cancer screening, in which 162,000 men were followed for 13 years to detect a decrease in the rate of death from prostate cancer of 1 per 10,000 person-years of follow-up!² It also illustrates that while screening tests can make a difference for individual patients, we also see their efficacy (or lack of it) played out on a population scale.

For more information on this topic – particularly regarding potential bias in screening – please see section III.20 (Evaluation of Screening Tests)

iii. **Some conflicts we may encounter when considering the use of screening tests**

Regardless of the value of any intervention, it is always useful to take a moment to consider some of the forces behind the curtain that drive the care that we deliver. In the case of screening tests, there are a few potential areas of conflict to highlight. As touched upon above, one conflict that clinicians may encounter with screening tests stems from the fact that we see the effect of screening tests primarily in population health statistics, but this information is applied in individual patient encounters.

For example, imagine a discussion with a patient in which you are addressing colorectal cancer screening. You read up ahead of time, and encountered a meta-analysis about flexible sigmoidoscopy in which there was a number needed to screen (NNS) of 450 to prevent one death from colorectal cancer.⁶ You feel conflicted about this, because that means the patient in front of you has *very* small chances from benefitting from this procedure – so even if there is a benefit of this screening intervention on a population scale, you find it hard to reconcile this with the numbers applied to your patient. While this type of conflict can occur with any intervention (anything with a NNT/S, in fact), it is more common with screening tests because they tend to have some of the largest NNTs around, and

hence benefit far fewer patients than is commonly acknowledged. We also see this in immunizations, which may explain some of the resistance encountered during the recent COVID19 pandemic.

Another point of conflict that some providers face is the use of screening rates as a performance measure. To continue the above scenario as our example, colorectal screening is incorporated into several measure sets for primary care practices.⁷ If this was happening in your practice, you would have an additional incentive to go ahead and send that patient described above for his flexible sigmoidoscopy. These measure sets (aka performance metrics) are not only used by systems as an evaluation tool for individual providers, but also may have implications for clinic funding – which often drive the delivery of care in resource-limited settings. Overall, performance measures can serve as a strong motivator for clinicians to increase their screening rates.⁸

iv. The evidence behind common screening tests

Now that we’ve reviewed the components of what makes effective screening tests and discussed some of the underlying conflicts, let’s dive into a specific example to get a better understanding of how all of this plays out in real-world examples. First of all, it may come as a surprise (or not if you are already familiar with screening tests) that not all of the screening tests we use are backed by well-designed randomized control trials. Because of the difficulty in acquiring the number of patients and follow-up to achieve adequate statistical power, some screening tests are backed by other types of studies (ex. prospective cohort studies), or use pragmatic trial designs which may compromise their validity. Often times, the guidelines we use in clinical practice from sources like the United States Preventative Task Force (USPTF) are based on a variety of studies (that may have used different follow-up lengths and age ranges) and systematic reviews – putting forth a best effort to synthesize information that is not always as satisfyingly clean as the well-designed RCT. The following section briefly describes some of the evidence behind screening mammography.

Breast Cancer – Screening Mammography (every 2 years)

As of the time of writing this (May 2023), the USPTF currently recommends biennial screening mammography for women ages 40-74 years old (Grade B Recommendation), which is based on a systematic review that the USPTF commissioned.³ When examining this systematic review, it combined data from several RCTs and reported mortality rates according to different age ranges. These data are provided in the table below with a calculated absolute risk reduction and number needed to screen for each age group.

Age Range	RR for breast cancer mortality (95% CI)	Reported Deaths Prevented per 10,000 women screened (95% CI)	Absolute Risk Reduction	Number Needed to Screen (NNS)
39-49 yrs old	0.92 (0.75 to 1.02)	2.9 (-0.6 to 8.9)	0.029%	3448
50-59 yrs old	0.86 (0.68 to 0.97)	7.7 (1.6 to 17.2)	0.077%	1299

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

60-69 yrs old	0.67 (0.54 to 0.83)	21.3 (10.7 to 31.7)	0.213%	469
70-74 yrs old	0.80 (0.51 to 1.28)	12.5 (-17.2 to 32.1)	0.125%	800

As you can see in the table above, the 95% confidence intervals cross zero for both the 39-49 and 70-74 age groups, and yet these age groups are included in the recommended screening population. To credit USPTF, their guideline does cite a need for additional research to evaluate the efficacy of screening these particular age groups.⁴

Also keep in mind that the numbers in the table above are for breast cancer-specific mortality; the systematic review reported that there was no difference (overall, or for any age group) for all-cause mortality between screening and control groups.³ This is not an uncommon feature of screening interventions and may reflect an element of overdiagnosis that is inevitable when we screen for disease: Consider the cancers that are slow growing and are not destined to cause clinically relevant disease – or even symptoms that are perceivable by a patient. When we detect this type of cancer, it causes inflation of disease incidence, leading to a decrease in disease-specific mortality: Using the ratio of deaths/cases, you can see that we would be adding patients to the denominator – patients who would otherwise not contribute to the mortality numbers.⁵ This is how cancer screening can simultaneously cause an increase in disease detection, a decrease in cancer-specific mortality among those screened, but also may not have an effect on overall mortality.

Estimating harms in screening is not an easy task, as many of the harms do not take the form of something as easy to track as a mortality rate. However, a meta-analysis of 3 RCTs from the U.K. investigating benefits and harms of breast cancer screening found that out of every 10,000 women screened (every 3 years from age 50-70), 129 would be over-diagnosed with breast cancer, and for every breast cancer death prevented, about three over-diagnosed cases would be identified and treated.⁵ False positive results and overdiagnosis can cause significant harms such as unnecessary biopsies, surgeries, and chemotherapy, but also contribute to significant psychological distress.

This is a complicated topic, but these harms are something that we should be aware of and counsel our patients about when it comes to screening tests. Be aware that many of these studies are focused on women at *average* risk for breast cancer. Perhaps a more nuanced approach needs to be developed to individualize recommendations for groups of women at low risk, and certainly for those at higher risk (e.g. women age 35-45, and those greater than 75).

v. **Summary and Conclusion**

Screening tests can be a tricky topic to communicate with patients, especially because the benefits we see from them are most evident on a population-scale, but individual patients have a low chance of benefitting from them because of large NNSs. Furthermore, the narrative in public media that screening “saves lives” often lacks a completeness of information: In this section, we’ve seen that

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

screening can be effective for decreasing disease-specific mortality, but may not have an effect on all-cause mortality. We also see harms from screening that include incidental findings, overdiagnosis, unnecessary treatments, false positive findings and psychologic distress. While in clinical practice we rely on the appraisal of evidence carried out by organizations such as USPTF, it is useful to examine some of the evidence yourself to be able to discuss this with patients in a more nuanced approach: As seen in the last section, some of the recommendations we use have difference efficacy for different patients, and always come with the risk of harms which are rarely discussed with patients ahead of time. Some patients may care about avoiding specific diseases (due to varying personal experiences), while others may only care about decreasing their all-cause mortality. Regardless of this, having an understanding of the evidence behind our screening tools may help you provide useful information to patients when making care decisions together.

References:

- 1) Herman CR, Gill HK, Eng J, Fajardo LL. Screening for preclinical disease: test and disease characteristics. *AJR Am J Roentgenol.* 2002 Oct;179(4):825-31. doi: 10.2214/ajr.179.4.1790825. PMID: 12239019.
- 2) Schröder FH, Hugosson J, Roobol MJ, Tammela TL, Zappa M, Nelen V, Kwiatkowski M, Lujan M, Määttänen L, Lilja H, Denis LJ, Recker F, Paez A, Bangma CH, Carlsson S, Puliti D, Villers A, Rebillard X, Hakama M, Stenman UH, Kujala P, Taari K, Aus G, Huber A, van der Kwast TH, van Schaik RH, de Koning HJ, Moss SM, Auvinen A; ERSPC Investigators. Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *Lancet.* 2014 Dec 6;384(9959):2027-35. doi: 10.1016/S0140-6736(14)60525-0. Epub 2014 Aug 6. PMID: 25108889; PMCID: PMC4427906.
- 3) Nelson, Heidi D et al. "Effectiveness of Breast Cancer Screening: Systematic Review and Meta-Analysis to Update the 2009 U.S. Preventive Services Task Force Recommendation." *Annals of internal medicine* 164.4 (2016): 244–255. Web.F
- 4) "Breast Cancer Screening." U.S. Preventive Services Task Force. Web. Accessed May 12, 2023: <https://www.uspreventiveservicestaskforce.org/uspstf/draft-recommendation/breast-cancer-screening-adults#citation48>
- 5) Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer.* 2013 Jun 11;108(11):2205-40. doi: 10.1038/bjc.2013.177. Epub 2013 Jun 6. PMID: 23744281; PMCID: PMC3693450.
- 6) Holme Ø, Bretthauer M, Fretheim A, Odgaard-Jensen J, Hoff G. Flexible sigmoidoscopy versus faecal occult blood testing for colorectal cancer screening in asymptomatic individuals. *Cochrane Database Syst Rev.* 2013 Oct 1;2013(9):CD009259. doi: 10.1002/14651858.CD009259.pub2. PMID: 24085634; PMCID: PMC9365065.
- 7) Health Plan Employer Data Information Set (HEDIS) sponsored by the National Committee for Quality Assurance (<http://www.ncqa.org/Programs/HEDIS/index.htm>)
- 8) Klabunde CN, Lanier D, Breslau ES, Zapka JG, Fletcher RH, Ransohoff DF, Winawer SJ. Improving colorectal cancer screening in primary care practice: innovative strategies and future directions. *J Gen Intern Med.* 2007 Aug;22(8):1195-205. doi: 10.1007/s11606-007-0231-3. Epub 2007 May 30. PMID: 17534688; PMCID: PMC2305744.

Submitted 5/17/2023

III.22 Challenges of Diagnostic Testing and Risk for Bias (Rebecca Robbins)

Clinicians rely on diagnostic testing to make decisions for and with their patients. Tests are ordered every day and decisions are made from these results. This process raises many questions. For example, how do we discriminate between tests and determine the accuracy of the results? How do we determine if a diagnostic test is valid? How is this done if there is no gold standard to which to compare the test? And how is bias introduced in diagnostic testing?

Basic Definitions in Diagnostic Testing

In order to understand diagnostic testing, we need to understand some basic definitions in statistics to help to help determine the diagnostic accuracy of a test. Diagnostic accuracy of any test helps us answer the question “how does this test discriminate between two conditions (health and disease)?” The principles that help us determine this include sensitivity and specificity, predictive values, likelihood ratios, area under the ROC (receiver operator curve) and Youden’s index. Below is a very brief overview of these principles, covered elsewhere in the Guide, but useful for one to refresh at this point.

The first step in calculating sensitivity and specificity is to make a 2 by 2 table with groups of subjects broken up in reference to the gold standard in the columns and categories according to the tests in rows.

Test	Subjects With Disease	Subjects Without Disease
Positive	TP	FP
Negative	FN	TN

Sensitivity- defines the proportion of true positive subjects with the disease in a total group of subjects with the disease ($TP/TP+FN$). This is the probability of a positive test result in subjects with the disease.¹

Specificity- proportion of subjects without the disease with a negative test result in total group of subjects without disease ($TN/TN+FP$). This represents the probability of a negative test in a subject without the disease.¹

Positive predictive value -the probability of having the disease of interest in a subject with a positive result ($TP/TP+FP$): Predictive values are largely dependent on disease prevalence in the population, unlike sensitivity and specificity, which are usually thought of as fixed characteristics of the test.¹

Likelihood ratio- the ratio of an expected test result in subjects with disease compared to the subjects without the disease. This is a very useful measure of diagnostic accuracy. For example, the LR of a positive test is the sensitivity divided by 1-specificity; for a negative likelihood ratio it is 1-sensitivity divided by the specificity. A positive LR tells us how many times more likely a positive test result is in subjects with disease compared to those without disease.¹

ROC curve- the shape of a ROC curve and the area under the curve (AUC) help to estimate how high is the discriminative power of a test. The area under the curve is a measure of diagnostic accuracy. The closer the curve is to the upper left hand corner and the larger area under the curve, the better the test is in distinguishing between diseased and non-diseased.¹

Diagnostic Odds Ratio- ratio of the odds of positivity in subjects with disease relative to the odds in subjects without disease. This is used for general estimates of discriminative power of diagnostic procedures and also for the comparison of diagnostic accuracies between two or more diagnostic tests. The equation is $DOR = (TP/FN)/(FP/TN)$.¹

Youden's index-One of the oldest measures of diagnostic accuracy. It is used for the evaluation of overall discriminative power and for comparing a test to another test. You subtract 1 from the sum of the test's sensitivity and specificity (sensitivity + specificity) – 1. For a test with poor diagnostic accuracy, Youden's index equals 0. For a perfect test, Youden's index equals 1.¹

Overall, these measures described above and measures of diagnostic accuracy as a whole, are very sensitive to the population and prevalence of a disease along with the spectrum of disease in the population. Additionally, the measures above are obtained by comparing the index test results, with the index test being the new test under evaluation, with the results of the best currently available test for diagnosing the disease. The test that is compared against the index test is called the reference standard, which in many instances is the gold standard. The gold standard would have a sensitivity and specificity equal to 100%, perfectly discriminating between subjects with and without the disease. However, in reality, it is not that simple. What do we do when there is no gold standard to which to compare a diagnostic test? Or what if the gold standard is an imperfect test, expensive or very invasive?

Absence of a Gold Standard

In a meta-analysis conducted in 2019, 6127 articles were identified and ultimately 209 articles were included in the review that helps to address the questions above when there is no gold standard or if the gold standard is prohibitively expensive or invasive.²

When the disease cannot be verified with a gold standard, imputation and bias correction methods can be used. This would include methods to correct for verification bias when the disease state of the subjects has not been verified. Verification bias will be discussed below. 48 different statistical methods were identified in this group. For subjects whose disease could not be verified with a gold

standard, another reference standard that was less accurate or less invasive was used. This is called differential verification. Three different statistical methods were identified to help with this approach, including a Bayesian latent class approach, a Bayesian method and a ROC approach. These methods can adjust for the differential verification bias and reference standard bias that results from using an imperfect reference standard.²

When using multiple imperfect reference standards, there are multiple methods that can be employed to help adjust for the imperfect reference standard. Discrepancy analysis can be used, which compares the index test with an imperfect reference standard. When there are discordant results, then subjects undergo another test, called a *resolver test*, to determine if they have a disease. However, this can lead to biased results. *Latent class analysis* can also be used by simultaneously using probabilistic models with the assumption that the disease status is latent or unobserved. Another method that can be employed involves constructing a composite reference standard. With this method, results from multiple imperfect tests are combined with a predetermined rule to construct a reference standard that is used to evaluate the index test. With the index test excluded as part of the composite reference standards, incorporation bias can be avoided. The last method is panel or consensus diagnosis, which uses the decision from a panel of experts to determine the disease status of each subject, which then can be used to evaluate the index test.² But you can see the challenges that are inherent to these remedies.

Bias in Diagnostic Testing

As seen above, bias can be a common problem seen with diagnostic testing. Bias can be introduced at all times during the study phases, including with patient selection, interpretation of the index test and in determining gold standards. Patient selection also clearly has a large impact on the diagnostic test characteristics. There are four main types of bias that can be a result of patient selection, including referral bias, spectrum bias through case control design, spectrum bias through dropping indeterminate subjects and spectrum bias through convenience sampling.³

Referral Bias

Referral bias, which is also known as partial verification bias, occurs when patients are selected based on either a positive or negative gold standard test. In these cases, patients are enrolled based on verification of the disease with the current gold standard test. This integrates bias into the test as a patient is more likely going to undergo the gold standard test if the index test is positive. This artificially increases true positives and increases sensitivity. To avoid this bias, the gold standard test should be performed on a random sample of patients with suspected disease, regardless of the result of the index test.³

Spectrum bias

Presentation of disease is not black and white and clinicians see a broad spectrum of diseases. This spectrum of disease needs to be incorporated into diagnostic testing. Spectrum bias can occur if the

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

spectrum of disease excludes ambiguous results or deviates from what is seen in clinical practice. Spectrum bias can be introduced in 3 different ways, including case control design, exclusion of indeterminate patients or convenience sampling.³

In spectrum bias due to case-control design, a group of patients known to have the disease (cases) and the group with no disease (controls) are given the index test. This can introduce bias by failing to have a group of patients representing the spectrum of disease. For example, in a group of patients who present to the ED with RUQ pain, investigators want to look at the specificity and sensitivity of a positive Murphy's sign in diagnosing cholecystitis. Their cases are patients with cholecystitis determined by pathology post surgery and their controls are patients who were discharged home. This misses a group of patients who were admitted to the hospital and found to have other diagnoses, where either pathology did not show an inflamed gallbladder or they were determined to have another cause of RUQ pain. This leads to a study where the controls are overall less sick.³

In spectrum bias due to dropping indeterminate subjects, bias is introduced by ignoring subjects with indeterminate results. Ultimately, dropping these results increases sensitivity and specificity by decreasing the denominator in the sensitivity and specificity equations.³

In spectrum bias from convenience sampling, patients could be left out of sampling for many reasons, such as difficulty in performing the test on them. This can result in a falsely increased sensitivity or specificity.³

Disease verification bias

Additionally, there can be disease verification bias, where the index test can either be incorporated into the gold standard or the gold standard is only applied to a certain population due to a variety of reasons, such as cost or invasiveness of the test. There can be *partial disease verification* and *differential verification bias*. *Partial verification bias* occurs when the subjects with a positive index test are more likely to receive the gold standard, which is discussed above. An example of this is when a patient with a positive EKG then undergoes a coronary catheterization. Then only those that receive the gold standard are included in the patient population. *Differential verification bias* occurs when more than one gold standard test is used and when these two gold standards classify the disease differently.⁴ This is also known as double gold standard bias.

Interpretation bias

Excluding indeterminate results from analysis can result in a spectrum bias, as noted above. If patients with indeterminate results are not excluded, it must be explicitly stated whether results are considered positive or negative in the analysis. An additional type of interpretation bias is due to review bias, where clinicians interpret tests based on prior information, leading to biased results. In studies of diagnostic tests, the interpreter of an index test is unblinded to whether the patient received the gold standard and thus, the gold standard's results. This may cause them to alter the

interpretation of the index test to agree with the gold standard results, ultimately falsely increasing the sensitivity and specificity of the index test.

How to Avoid Bias in Diagnostic Testing?

As seen above, there are many areas where bias can occur in using diagnostic testing. Clinicians can only determine the risk of bias results if the necessary information is provided to them. The Standards for Reporting of Diagnostic Accuracy Studies Guidelines were created to help with the reporting of methods in studies of diagnostic tests and to help create more transparency. It was initially released in 2003 and then updated in 2015. The updated STARD list now has 30 essential items, including key points to include in an abstract, intro, methods, results and discussion.⁵ Additionally, the QUADAS tool was developed to assess the quality of diagnostic accuracy studies. It consists of 4 key domains that discuss patient selection, index test, reference standard and flow of patients through the study and timing of the index tests and reference standard. An updated version also rates risk of bias and concerns about applicability and discusses handling studies in which the reference standard consists of follow up.⁶ Both of these guidelines are ultimately important in avoiding bias, or at least recognizing bias, in diagnostic testing.

References

1. Šimundić AM. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC*. 2009;19(4):203-211. Published 2009 Jan 20.
2. Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard - An update. *PLoS One*. 2019;14(10):e0223832. Published 2019 Oct 11. doi:10.1371/journal.pone.0223832
3. Hall MK, Kea B, Wang R. Recognising Bias in Studies of Diagnostic Tests Part 1: Patient Selection. *Emerg Med J*. 2019;36(7):431-434. doi:10.1136/emered-2019-208446
4. Kea B, Hall MK, Wang R. Recognising bias in studies of diagnostic tests part 2: interpreting and verifying the index test. *Emerg Med J*. 2019;36(8):501-505. doi:10.1136/emered-2019-208447
5. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HC, Bossuyt PM. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016 Nov 14;6(11):e012799
6. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011 Oct 18;155(8):529-36.

Submitted 2/2022

III.23 Bayes Theorem (Malachy Sullivan and Alex Briand)

Bayes' theorem provides a way in which new information affects the likelihood of an event or outcome as when a test is applied in a clinical scenario.

Bayes' theorem states, in common parlance, that the probability of an event (i.e., a diagnosis) depends on new information (results of a diagnostic test) applied to what is previously known about an event (pre-test probability).

To put this into a clinicians' terms, Bayes' theorem helps us once we assign a pretest odds of a patient having a specific diagnosis to then predict how a positive or negative test will impact the post-test odds. While most of us are already inherently thinking in a similar manner to this, it can be useful to be explicit about and to delve into the mathematical equations to fully understand this concept.

Let's look at some definitions first.

Pre-test probability = chance of event occurring/all events x100

Pre-test odds = pretest probability/1-pretest probability (or in other words, the probability of event happening/probability of event NOT happening)

Likelihood ratios (LR) are derivations of sensitivity and specificity (inherent characteristics of tests), and are expressed as:

Likelihood ratio + (or LR+) = sensitivity/1-specificity

Likelihood ratio - (or LR-) = 1-sensitivity/specificity

Post-test odds = Likelihood ratio (positive or negative) x pretest odds

To summarize, probability is the chance an event occurs compared to ALL events, while odds equals the chance an event occurs compared to the chance the event DOES NOT occur. This is a subtle difference, but as clinicians we do not generally think of a patient as either having x disease versus all other diseases because this list can be quite extensive, but rather we think of the probability of having x disease versus *not having the disease (i.e., the odds)*.

So now that we have clarified how to convert a pretest probability into pretest odds, you might still be wondering how this relates to likelihood ratios. Sensitivity is often called your "true positive rate". Specificity is your "true negative rate." 1-specificity is "false positive rate" and 1-sensitivity gives you a "negative rate." A positive likelihood ratio compares the odds that a positive test is a true positive versus the chance it is a false positive. Likelihood ratios are essentially the odds that a positive or negative test result is a true result. A common pitfall is to think about sensitivity and specificity in a vacuum, but we can see from the above equations that just because a test has a very high sensitivity it can be so nonspecific that a positive test does not actually change your post-test probability (ie think ESR in a patient with fever).

Here are some more relationships that can be derived by looking at pretest probability, Sn and Sp.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Let p = prior probability of disease

$(1-p)$ = prior probability of no disease

Se = test sensitivity

Sp = test specificity

As Table 1 indicates,

PPV = $p(\text{Se})/[p(\text{Se}) + (1-p)(1-\text{Sp})]$, and

Post-test odds = $p(\text{Se})/[(1-p)(1-\text{Sp})]$.

To reach a statement of PPV as a function of post-test odds, the algebraic trick is to take the inverse of the equations for both values; thus

$1/\text{PPV} = [p(\text{Se}) + (1-p)(1-\text{Sp})]/p(\text{Se}) = 1 + (1-p)(1-\text{Sp})/p(\text{Se})$, and $1/\text{post-test odds} = (1-p)(1-\text{Sp})/p(\text{Se})$.

We can make use of the presence of $(1-p)(1-\text{Sp})/p(\text{Se})$ in both equations to say that

$(1/\text{PPV}) - 1 = 1/\text{post-test odds}$, and

$1/\text{PPV} = (1/\text{post-test odds}) + 1 = (1/\text{post-test odds}) + (\text{post-test odds}/\text{post-test odds}) = (1 + \text{post-test odds})/\text{post-test odds}$.

Taking the inverse of both sides,

$\text{PPV} = (\text{post-test odds})/(1 + \text{post-test odds})$.

Let's look at two examples:

Number with disease: 20		Number without disease: 80	
Disease +		Disease -	
Test +	(a) 18	(b) 8	
Test -	(c) 2	(d) 72	

First, specify a pre-test (prior) probability of the disease in question. Second, assign values of sensitivity and specificity to the test in question, using decimals. Third, imagine a group of 100 patients identical to the patient in question. With the estimate of prior probability, calculate how many do and not have disease. Put these two numbers over the columns of a 2 x 2 table. Using the designated sensitivity and specificity of the test in question, compute the number that belongs in each cell of the table. Then calculate the desired post-test parameter(s).

Recall this from an earlier chapter:

Sensitivity and Specificity*

Result of Test Investigated	Result of Gold Standard Test	
	Disease Positive	Disease Negative
Positive (+)	TP (a)	FP (b)
Negative (-)	FN (c)	TN(d)

TP= True positive
 FP= False positive
 TN= True negative
 FN= False negative

Now consider the following:

Assume that the prior probability of disease is 0.2, and the sensitivity and specificity of the test to be performed are both 0.9. The table describing results in 100 identical patients is as follows:

With a positive test result, the pre-test (prior) probability of disease, 0.2, is converted to the post-test (posterior) probability, 18/26, or 0.69. The pre-test odds of disease, 0.25, are converted to the post-test odds, 18/8, or 2.25. (Note that pre-test odds, 0.25, x the likelihood ratio, 0.9/0.1 or 9, also equals 2.25.)

Let us take a look at another clinical example.

An 18-year-old male presents with abdominal pain. You wish to evaluate the possibility of appendicitis in this patient. If your determined pre-test probability is 20% (1 in 5 patients with this presentation have appendicitis) you may opt for a CT scan to further evaluate. A CT scan in adults with appendicitis has a sensitivity of 95% and a specificity of 96%, giving it a positive LR of 18.8 and a negative LR of 0.06.

Therefore:

To begin, we convert our pre-test probability into pre-test odds: $0.2 / (1-0.2) = 0.25$ CT Scan positive: pre-test odds of $0.25 * 18.8 = 4.7$.

Our post-test odds are 4.7. We now convert this to a post-test probability: $4.7 / (1+4.7) = 82.4\%$
 82.4% post-test probability of appendicitis

Now, if the scan is negative:

CT Scan negative: pre-test odds of $0.25 * 0.06 = 0.015/1.015 = 1.5\%$ post-test probability of appendicitis

As you can see in the above example, the pre-test odds do have a significant impact on our post-test probability, even in a diagnostic test as accurate as a CT scan. If our patient presented with nausea, abdominal pain, and vomiting after eating old leftovers and has several sick family members, our pre-test probability of appendicitis may be more like 2% (1 in 50 patients with this presentation have appendicitis). Let's run through Bayes' theorem with this data:

Pre-test probability to pre-test odds: $0.02 / (1-0.02) = 0.02$
 CT Scan positive: pre-test odds of $0.02 * 18.8 = 0.376$
 $0.376 / (1 + 0.376) = 27\%$ post-test probability of appendicitis.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

CT Scan negative: pre-test odds of $0.02 * 0.06 = 0.0012 = 0.0012 / (1+0.0012) = 0.1\%$ post-test probability of appendicitis.

As you can see, such a patient would likely not benefit from a CT scan to specifically evaluate appendicitis as even a positive test still has a low post-test probability. The ultimate question is, how does one obtain an accurate pre-test probability? Through clinical application of your history, exam, and utilization of Bayes' theorem in clinical practice. A patient presenting with chest pain will have a different pre-test probability of MI if their pain is reproducible or if they describe their pain as 'crushing' and associated with arm pain. While a clinician may not actually do the math as we did above, their clinical gestalt utilizes Bayes' theorem and assists them in guiding future work-up.

As you also see, the best use of testing is in the mid-range of prior probabilities; if the PP is extremely low, even a test with a good LR+ will not move your post-test probability high enough for approaching the certainty you are looking for, while when the PP is very high, a test with a good LR- will not move the posttest probability low enough to justify not treating the patient.

Some final thoughts:

Regarding Bayes' Theorem: accurate test interpretation is not the clinician's ultimate goal. This skill is a tool. The real goal is to determine -- before doing the test -- whether and how a positive or negative result will affect further testing or treatment. For example, in the abdominal pain vignette above, the pre-test probability of appendicitis is 0.2; the post-test probability is 0.86 after a positive result and 0.01 after a negative result. Surgery will be performed after a positive result and withheld after a negative result. However, one can argue that surgery should be performed if the probability of appendicitis is ≥ 0.15 . Although the revision of that probability is greater with a positive than with a negative test, the negative result appears more likely to influence therapeutic decision-making.

In the other vignette, the prior probability of appendicitis is 0.02 because the patient has a second disorder that can explain the clinical manifestations. A negative CT scan revises the prior probability further downward and does not affect therapeutic decision-making. A positive scan raises the probability of appendicitis to 0.27, a value that may justify surgery.

In both vignettes, a strong case can be made for doing the CT scan because one of the two possible results will guide treatment. All possible contingencies can and should be articulated in advance of the test. In general, this exercise necessitates consideration of risks associated with the test, facilitates clearer communication of rationale with patients and families, and invites timely consultation with colleagues who may be involved in treatment.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Regarding specification of pre-test probability of disease: Ideally, the clinician identifies *every possible disease* in the differential diagnosis and assigns each one a pre-test probability. The sum of all of the pre-test probabilities is 1. The following logical error occurs with some frequency: Say that a presentation occurs which is not is not typical of any disease. Consultants flock to the patient and proclaim that he has no condition in their areas of expertise. They assert that the diagnoses in which they are expert are *unlikely* or even impossible because the patient's presentation is so *atypical* of those conditions. Indeed, the logical conclusion from the aggregate of consultations may be that the patient has no disease at all. However, *likelihood* must be distinguished from *typicality*. If four conditions are under consideration, and the presentation is extremely but equally atypical of all four, each condition has a pre-test probability of 0.25.

All of this together should highlight some key characteristics of a thoughtful clinician. Based upon a patient's presenting symptoms, exam, vitals, and available diagnostics, what is their current probability of having a confident (i.e., likely) diagnosis? Having a predefined threshold for diagnosis and initiating treatment certainly might change depending on the gravity of the situation (i.e., treating a cellulitis with antibiotics versus PLEX for TTP). This all relies on your interpretation of the available data and will inevitably vary with others' assessments from time to time based on their prior experience and knowledge level. If you have not reached a threshold for diagnosis, you may be hesitant to start treatment, and you will most likely need to obtain additional testing. Knowing the test characteristics of whatever you order (i.e., Sn/Sp) and knowing how a positive or negative result will impact your decision making is crucial to picking the right test.

So have fun with these valuable concepts and practice their application in your daily work!

Sept 2016 (Malachy) and Jan 2020 (Alex)

(This section was edited with contributions from Dr. K.R. Phelps, Albany Medical College, who kindly reviewed this material. Jan 2018)

III.24 Beyond Bayes: Some Issues in Diagnostic Reasoning; or, Towards an Evidence-based Framework for Diagnostic Reasoning (Stephen Conn GSM4)

In a 1959 paper in *Science*, Ledley and Lusted introduced the notion of Bayesian reasoning in diagnosis in clinical medicine. They spend some time developing a logical/symbolic formalism for diagnostic reasoning, and conclude by remarking:

The "most likely" diagnosis is determined by calculating the conditional probability that a patient presenting these symptoms has each of the possible disease complexes under consideration. This probability depends upon two contributing factors. The first factor is the conditional probability that a patient with a certain disease complex will have a particular symptom complex; it remains relatively independent of local factors and depends primarily on the physiopathological effects of the disease complex itself. The second factor is the effect on medical diagnosis of the circumstances surrounding the patient or, more precisely, the total probability that *any* person chosen from the particular population sample under consideration will have the particular disease complex under consideration; this may depend on the geographical location of the population sample, or the season when the sample is chosen, or whether the population sample is chosen during an epidemic, or whether the sample is composed of patients visiting a particular type of specialist or clinic, and so forth.

Simplifying for sake of brevity, an approximation of their argument is the common distillation that learners in medicine are taught today: In order to deploy a diagnostic test in clinical reasoning, or to understand how a physical exam finding impacts the likelihood of a diagnosis, one should consider the prevalence of the disease under suspicion to inform the prior probability that the patient has the disease; then, using Bayes' formula and clinical information about the patient's presentation, or data on the performance characteristics of the diagnostic test, one can arrive at the posterior probability that the patient does or does not have the disease under consideration. (See previous chapter on Bayes theorem as well.)

Today, diagnostic reasoning remains a difficult topic to teach, with organizations dedicated to improving it, such as the Society to Improve Diagnosis in Medicine, and with many authors noting the challenges of effectively teaching medical learners how to deploy concepts such as sensitivity/specificity, positive and negative predictive value, and likelihood ratios in thinking diagnostically. Further, a large body of literature examines the frequency of diagnostic error in various domains of medicine, and diagnostic error is thought to be a major source of adverse events in the health system, including iatrogenic deaths (Raffel et al., 2020). Some authors have proposed deploying tools to identify possible diagnostic errors, such as the Safer Dx instrument, and have attempted to validate the ability of these tools to identify errors (Al-Mutairi et al., 2016).

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Simultaneously, a relatively small number of authors since Ledley and Lusted have highlighted concerns with or potential flaws in their proposed framework for diagnostic reasoning. Miettinen and Caro noted in a 1992 paper that the “junction” at which Bayes’ formula is applied is not well-defined in the original paper, and varies significantly across clinical practice: for Ledley and Lusted, the aggregate of “non-clinical” data not able to be informed by “medical knowledge” is what informed the prior probability, and the totality of the clinical data and *a priori* knowledge about known relationships between constellations of symptoms and complexes of diseases is what allowed a diagnostician, via the application of Bayes’ formula, to arrive at the probability of various diagnoses. Miettinen and Caro correctly note that this approach diverges from what we are often taught today, when the totality of information about the patient may be considered to inform a ‘pre-test’ probability of a disease, and then a diagnostic test result and some known data on the performance of the test can then be used to derive a ‘post-test’ probability. Alternately, they also note the approach to sequentially apply “Bayesian” reasoning in the attempts to account for each of the successive facts about a patient’s presentation, generating a long series of prior and posterior probabilities, informs the consideration of the next fact, .

For each of these “junctions,” they highlight serious formal or practical issues with the Bayesian project. Highlighting one example, the authors note that the referent group from which a prior probability of a disease is often difficult to define rigorously. In today’s terms, suppose a 50-year-old man presents to Massachusetts General Hospital with a cough, and a diagnostician has concern for COVID-19. Learners are taught that the Bayesian paradigm requires a rigorous diagnostician to consider the prior probability that the patient has COVID-19 in helping them to interpret the results of any diagnostic test. A casual questioner might ask, “What is the current prevalence of COVID-19 in Boston around the time of the patient’s presentation?” A deeper observer might ask, however, whether the patient’s age, gender, or chief concern of cough should inform the prior probability. What about the hospital to which the patient presents—is it correct to consider what is known about the prevalence of the disease in Boston, in Massachusetts, in the patient’s zip code, in the specific hospital to which he presents? What of a past medical history of asthma? What about the season of the patient’s presentation and known facts about seasonality of transmission of disease? The list of potential considerations is endless, adding to the complexity of inputs that inform the prior probability.

A practical observer might retort that despite these nooks and crannies, the *result* of the diagnostic reasoning will surely be superior when a diagnostician makes an attempt to consider at minimum some *approximate* prevalence of the disease to inform interpretation of a diagnostic test. For example, even broad lower and upper bounds on the prevalence of COVID-19 could be used to generate drastically different positive and negative predictive values for the interpretation of a COVID-19 test when compared to a diagnostician considering only what is known about the sensitivity and specificity of the test.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Other authors have raised additional thorny issues regarding the nuts and bolts of our current framework. Sensitivity and specificity are often assumed to be fixed characteristics of a test, independent of disease prevalence. Some have asserted that this assumption is not precisely correct, measuring variations in sensitivity and specificity of common diagnostic tests across different populations or subgroups who may have a given disease (Ransohoff and Feinstein, 1978; Lachs et al., 1992; Goehring et al., 2004). To some extent, these objections parallel those that Miettinen and Caro made to the original application of Bayes' formula—how can a diagnostician be sure that the individual patient before them is “similar enough” in characteristics, whether demographic or clinical, to the population of the study establishing a diagnostic test's accuracy in order to be confident that the test will perform as expected? Variation in test performance across subgroups is common and is now termed the “spectrum effect”, and when the spectrum effect alters positive or negative predictive values, this is called “spectrum bias” (Goehring et al., 2004). The extent to which spectrum bias meaningfully impacts clinical decision making is not well understood, although Goehring et al. do think through a few specific examples from other studies and offer some generalizable principles for clinicians' awareness.

Others have pointed out that while new diagnostic tests are compared to “gold standard” reference tests which are assumed to be perfect, the gold standard tests themselves are not necessarily so, and this can impact how we understand the performance of new diagnostic tests (Boyko et al., 1988).

In the more than sixty years since the Bayesian framework was introduced, much has changed about the practice of medicine, including the introduction of evidence-based medicine (EBM), the subject of this course. In its most abstract, EBM demands that clinical decision-making be supported by robust data. In the plane of decision-making regarding therapy for disease, most often this means that randomized clinical trials demonstrate that a therapy is superior to placebo or a preceding standard of care before it be recommended by a new clinician or adopted as the new standard. Often, we prefer that these trials be performed across multiple centers or are reproduced by multiple authors; we also can deploy strict requirements about the outcomes that a therapy must impact or improve in order to justify its recommendation. We don't always hold to these standards in recommending or selecting therapies, but in many specialties of medicine we at least aspire or work towards them.

No such framework exists for diagnostic decision making. Is it possible or realistic to envision a future in which diagnostic reasoning is held to the same standards as therapeutic decision making? Diagnostic tests should be held to as close to the same evidence standards as treatments as possible: Does a diagnostic test benefit a given patient? What is the patient-centered outcome that performing the diagnostic test will improve?

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Unfortunately, the areas in which we have this kind of evidence are very limited. Some commonly performed diagnostic tests and “pathways” have been examined in the literature, including, for example, abdominal CT imaging in the emergent evaluation of the acute abdomen (Mills et al., 2015). The concern expressed in many of these studies is around overuse of diagnostic tests, particularly imaging, and emphasizing how these studies do not improve diagnostic accuracy or patient-centered outcomes. For some narrow domains, “clinical decision rules” attempting to codify when certain diagnostic imaging should be used have been developed and validated; however, Mills et al. note that deploying these CDRs has run into “practical challenges” and their uptake so far appears limited. Thus we are still early in applying the rigor and clarity that diagnostic reasoning demands, for learners, clinicians, patients and policy makers. It’s quite an invitation for us all!

References:

1. Al-Mutairi, Aymer, Ashley N. D. Meyer, Eric J. Thomas, Jason M. Etchegaray, Kevin M. Roy, Maria Caridad Davalos, Shazia Sheikh, and Hardeep Singh. 2016. “Accuracy of the Safer Dx Instrument to Identify Diagnostic Errors in Primary Care.” *Journal of General Internal Medicine* 31 (6): 602–8.
2. Boyko, E. J., B. W. Alderman, and A. E. Baron. 1988. “Reference Test Errors Bias the Evaluation of Diagnostic Tests for Ischemic Heart Disease.” *Journal of General Internal Medicine* 3 (5): 476–81.
3. Goehring, Catherine, Arnaud Perrier, and Alfredo Morabia. 2004. “Spectrum Bias: A Quantitative and Graphical Analysis of the Variability of Medical Diagnostic Test Performance.” *Statistics in Medicine* 23 (1): 125–35.
4. Lachs, M. S., I. Nachamkin, P. H. Edelstein, J. Goldman, A. R. Feinstein, and J. S. Schwartz. 1992. “Spectrum Bias in the Evaluation of Diagnostic Tests: Lessons from the Rapid Dipstick Test for Urinary Tract Infection.” *Annals of Internal Medicine* 117 (2): 135–40.
5. Ledley, R. S., and L. B. Lusted. 1959. “Reasoning Foundations of Medical Diagnosis; Symbolic Logic, Probability, and Value Theory Aid Our Understanding of How Physicians Reason.” *Science* 130 (3366): 9–21.
6. Miettinen, O. S., and J. J. Caro. 1994. “Foundations of Medical Diagnosis: What Actually Are the Parameters Involved in Bayes’ Theorem?” *Statistics in Medicine* 13 (3): 201–9; discussion: 211–15.
7. Mills, Angela M., Ali S. Raja, and Jennifer R. Marin. 2015. “Optimizing Diagnostic Imaging in the Emergency Department.” *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine* 22 (5): 625–31.
8. Raffel, Katie E., Molly A. Kantor, Peter Barish, Armond Esmaili, Hana Lim, Feifei Xue, and Sumant R. Ranji. 2020. “Prevalence and Characterisation of Diagnostic Error among 7-Day All-Cause Hospital Medicine Readmissions: A Retrospective Cohort Study.” *BMJ Quality & Safety* 29 (12): 971–79.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

9. Ransohoff, D. F., and A. R. Feinstein. 1978. "Problems of Spectrum and Bias in Evaluating the Efficacy of Diagnostic Tests." *The New England Journal of Medicine* 299 (17): 926–30.
10. Singh, Hardeep, Arushi Khanna, Christiane Spitzmueller, and Ashley N. D. Meyer. 2019. "Recommendations for Using the Revised Safer Dx Instrument to Help Measure and Improve Diagnostic Safety." *Diagnosis (Berlin, Germany)* 6 (4): 315–23.

Submitted Nov 2021

III.25 The Placebo Effect- should we pay attention? (Devin van Dyke, GSM4)

The role of evidence in clinical practice is to inform decision-making. When patients or families are involved in the decision-making process, as they should almost always be, the translation of research evidence into an understandable form for patients is a central task of the clinician. Synthesizing knowledge of the patient and knowledge of the evidence in order to deliver useful information and advice is in some sense the essential role of the physician, and even the pinnacle of EBM organization, the “system” that incorporates chart data with evidence to provide personalized recommendations, requires a skillful human touch to apply these to the patient in a way that fits their needs and goals.

One of the most useful patient-friendly ways of presenting research findings is the absolute risk difference (ARD), or its closely related sibling, the Number Needed to Treat, or NNT. The NNT can be further refined into either a number needed to benefit (the number of patients who must be treated in order for one to benefit) or number needed to harm (the number of patients who when treated will result in one who is harmed). These measures show the benefit (or harm) of the intervention, taking into account the prevalence of the outcome in the population of interest in the absence of intervention (ARD), or, alternately, provide an estimate of the number of patients who must receive an intervention in order to prevent (or potentially to cause) one outcome of interest (NNT). These are helpful measures for talking to patients and can often be understood without too much difficulty. These measures are computed on the basis of findings of RCTs, which typically compare an active intervention with a placebo intervention, although there are many RCTs comparing active treatments to demonstrate superiority or inferiority. This study design is ideal for determining the benefit of the intervention, which is not contingent on the placebo effect, but when counseling patients it may potentially lead to distortion of the evidence used in decision-making. The reason for this is that patients who receive interventions in the clinic may benefit from the intervention itself as well as from the placebo effect. Theoretically, they may experience an even greater placebo effect because patients often know in the context of research studies that they may receive a placebo whereas in the clinic patients know that the interventions they receive are always active (and ideally evidence-based).

This failure to account for the benefits of the placebo effect leads to absolute risk differences that are probably lower and NNTs that are probably higher than the true values that groups of patients are likely to experience. Factoring considerations about placebo effects into the usual conversation one might have introduce more complexity into shared decision making. (Note that when saying to a patient, “The NNT for this intervention is 10, which means that I would have to apply this intervention to 10 patients in order for one to avoid the outcome for which I am treating you. We cannot know that you are going to benefit, but that 10% of patients so treated will benefit, and 90% will not.”). How can this distortion in the evidence and obstacle to accurate counseling be resolved?

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Researchers have attempted to quantify the size of the placebo effect. This is a complex effort as the placebo effect is highly dependent on many factors in study design (e.g., could the placebo be convincingly mistaken for the active intervention?) and appears to differ greatly based on the type of illness and outcome measure in question (e.g., processes involving immunity, the autonomic nervous systems, or mental states such as anxiety and pain are more susceptible to placebo whereas hyperacute processes like heart attacks, degenerative diseases, and hereditary diseases are less susceptible). A meta-analysis of studies with active, placebo, and no treatment arms conducted in 2001 by Hróbjartsson and Gøtzsche did not find evidence of a clinically relevant placebo effect, but a re-analysis of the dataset published in 2005 by Wampold generated a framework that we may use to guide patient counseling efforts. Wampold's framework classifies studies along two domains, analogous to those introduced above. The first is amenability of the disease process to psychological factors, from not amenable (anemia, bacterial infection), to possibly amenable (acute pain, chemotherapy-induced nausea, asthma), to definitely amenable (insomnia, chronic pain, depression). The second is adequacy of the study design to produce a robust placebo effect, with studies classified as 'adequate' if they were double-blinded, study participants were aware that they could receive a placebo and were aware when it was administered, and the treatment and the placebo were indistinguishable. This framework thus sorts studies into one of six categories, and despite some imperfections (e.g., there are probably more shades of adequacy than are accounted for here, the possibility of detection via side effects is not accounted for etc.) it provides a useful heuristic.

In the Wampold group's reanalysis of the 114 studies (involving 8525 patients) included in Hróbjartsson and Gøtzsche's 2001 meta-analysis, they found significant placebo effects for continuous outcomes in studies with adequate designs and definitely amenable disease processes, with an effect size of 0.29 (95% CI: 0.06 to 0.52). Notably, this effect size is comparable to the active treatment effect size of 0.24 (0.00 to 0.47), and no difference was seen between placebo effect sizes when subjective vs objective continuous outcomes were used. Adequately designed studies with possibly amenable or not amenable disease processes did not show a significant placebo effect, with Cohen's *d* of 0.17 (95% CI: -0.01 to 0.36) and -0.03, respectively. When analyzing the adequately designed studies with dichotomous outcomes, no significant placebo effect was found; the authors note that no significant effect was found for the active treatment either, which precludes the finding of a significant placebo effect as the placebo effect is unlikely to exceed the active treatment effect.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Wampold's group concluded that while power was limited by the small number of trials in each of their categories, there was evidence that the placebo effect had a meaningful benefit in amenable disease processes, and that this effect was comparable to the effect of many active interventions. Using the Kraemer method, we can convert the placebo effect size in definitely amenable disease processes (0.29) to a NNT, which comes to approximately 7. While we are limited by the relatively small number of studies analyzed here and risk of bias that is highlighted by Hróbjartsson and Gøtzsche in their rebuttal, and further study in this area is warranted, this conclusion can inform a shared decision-making conversation. Patients should be counseled that when the disease process is amenable to psychological influence, they are likely to benefit from an intervention to a greater degree than the ARD and NNT would indicate, perhaps as much as twice that degree, if the placebo effect is truly comparable to the active treatment effect in selected disease processes. While the Kraemer method has weaknesses (the Furukawa method is superior but requires estimation of the control event rate), a NNT of 7 represents an important influence on patient health that should not be ignored. Though it is a complicated topic to address and further research is required, clearly discussion of the placebo effect should find its way into shared decision-making in the clinic, and it is hoped that this article provides some assistance in this important effort.

References:

1. Hróbjartsson, A., & Gøtzsche, P. C. (2001). Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *New England Journal of Medicine*, 344, 1594–1602
2. Hróbjartsson, A., & Gøtzsche, P. C. (2007). Powerful spin on conclusion in Wampold and colleagues' re-analysis of placebo vs. no-treatment trials despite similar results as in original review. *Journal of Clinical Psychology*, 63, 373–377.
3. Hunsley J, Westmacott R. Interpreting the magnitude of the placebo effect: mountain or Molehill? *J Clin Psychol*. 2007 Apr;63(4):391-9. doi: 10.1002/jclp.20352. PMID: 17279525.
4. Furukawa TA, Leucht S. How to obtain NNT from Cohen's d: comparison of two methods. *PLoS One*. 2011 Apr 27;6(4):e19070. doi: 10.1371/journal.pone.0019070. PMID: 21556361; PMCID: PMC3083419.
5. Wampold, B. E., Minami, T., Tierney, S. C., Baskin, T. W., & Bhati, K. S. (2005). The placebo is powerful: Estimating placebo effects in medicine and psychotherapy from randomized clinical trials. *Journal of Clinical Psychology*, 6, 835–854.
6. Wampold BE, Imel ZE, Minami T. The placebo effect: "relatively large" and "robust" enough to survive another assault. *J Clin Psychol*. 2007 Apr;63(4):401-3; discussion 405-8. doi: 10.1002/jclp.20350. PMID: 17279522.

Submitted November 2020

III.26 The Big Data Paradox: A Conundrum of Abundance and Accuracy (Maria Malik GSM4)

“All models are wrong, but some are useful.” – G.E.P. Box

INTRODUCTION

In the realm of medical and healthcare research, the use of large data sets with a vast number of patients often is touted as the most effective way to attain accurate results. One might assume that this enables a sharper and clearer picture of whether a drug can really improve patient outcomes or whether a risk factor for a disease is legitimately associated with it. After all, the Law of Large Numbers theorem implies that as a sample size increases, the mean of the sample will be closer to the true mean of the population. Yet, studies with increasingly large sample sizes are also more vulnerable to heavy bias and misleadingly narrow confidence intervals – a phenomenon known as the “big data paradox.”

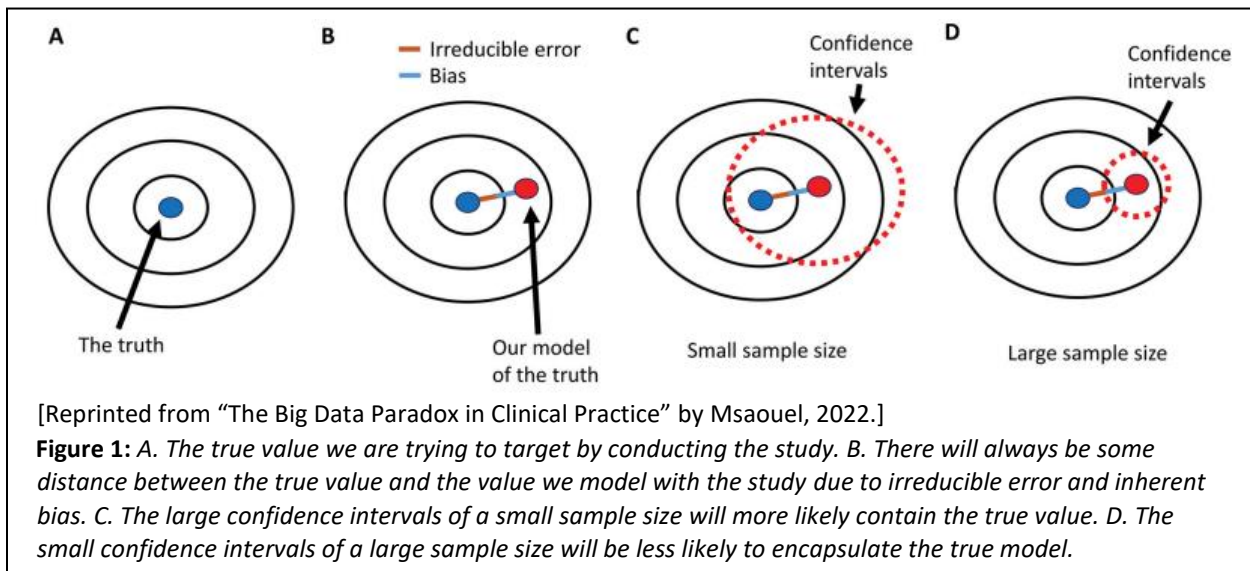
Table 1: Definitions of Key Statistical Concepts Discussed in this Chapter	
Bias	The distance between the true value of the parameter and the computed value based on a statistical model. It occurs due to the limitations and imposed assumptions of the statistical model.
Variance	A measure of dispersion – how spread out a set of values are. It is the standard deviation squared.
Standard Deviation	A measure of dispersion – how far apart values are in a data set relative to their mean. It is the square root of variance.
Confidence Interval	The probability that a parameter will fall between a set of values a certain proportion of times (i.e., 95% or 99% of times). If the confidence interval contains the null hypothesis, one cannot rule out that the noted observation is due to chance.
Standard Error	A measure of accuracy – how different the overall population mean is likely to be from the sample mean based on the standard deviation of the sample.
Random Error	Coincidental errors caused by unknown and unpredictable changes in an experiment.
Systematic Error	Consistent errors caused by flaws from a measuring instrument or due to other causes that produce a repeating error that can be fixed or proportional.
Reducible Error	Errors that can be removed to improve a model. They occur due to a combination of random error and systematic error.
Irreducible Error	Errors that are inherent to a model and cannot be removed. They occur due to unknown elements that are not represented within a dataset.

THE BIG DATA PARADOX DEFINED

In the big data paradox, as the sample size of a study increases, the probability that the confidence intervals obtained from that study will include the true value decreases (Msaouel, 2022). To further elucidate this concept, we can go back to basic theoretical statistics: as the sample size of a study increases, variance decreases. As variance decreases, standard error

also decreases. Thus, the confidence intervals for the results of this study become narrower. This inverse relationship of the width of confidence intervals and the sample size can be illustrated by two similar trials studying the impact of pembrolizumab versus placebo on the survival rate of patients with non-small-cell lung cancer: KEYNOTE-024 enrolled 305 patients (with primary survival rate endpoint confidence interval of 0.41 to 0.89) while KEYNOTE-189 enrolled over double that amount with 616 patients (with primary survival rate endpoint confidence interval of 0.38 to 0.64) – the difference in length of confidence intervals is 0.48 vs 0.26 (Gandhi et al., 2018; Reck et al., 2016).

With narrower confidence intervals, the probability that the true values lie outside that range increases (**Figure 1**). Conversely, the wider confidence intervals of a smaller study will have a higher probability of containing the true value. This big data paradox can be observed in all types of trials, including randomized control trials (RCTS).



The unreliable predictions made during the 2016 U.S. presidential election have been linked to the big data paradox as miniscule data defects in self-reported data compounded to result in estimates that widely underestimated one candidate’s vote share (Meng, 2018). During their in-depth statistical analysis of the estimation follies that could have led to the misleading predictions in 2016, Meng et al. (2018) used theoretical models to prove that on average, the larger a state’s voter population, the further away the actual voting share was from the typical 95% confidence intervals.

Essentially, in having a larger sample size, studies can be prone to using more simplified models, which may give way to uncertain assumptions about the population or not be able to capture participant heterogeneity (Msaouel, 2022; Senn, 2022). The trade-off for the low level of variance in such studies is that there is a higher risk for bias and systematic error (Msaouel, 2022). Nonetheless, while it would be easier to identify and account for bias in smaller trials and their larger confidence intervals are more likely expected to hold the true values that are

trying to be ascertained, they might be less useful in clinical practice because the individuals included in the sample might be widely divergent. Fewer datapoints which may be internally highly divergent may not yield results that are as generalizable in making decisions about patient care. Therefore, in thinking about how to conduct studies that can yield results applicable to clinical practice, it makes sense to have larger sample sizes with reduced variance, albeit systematic error and bias may be larger.

In clinical research, the overabundance of healthcare data as well as the desire to enroll a large number of patients for clinical trials mean that the big data paradox should be considered when interpreting research data and results. In the rest of this chapter, the mechanisms that may underlie this phenomenon and the strategies that can be used to mitigate it are reviewed.

MECHANISMS THAT UNDERLIE THE BIG DATA PARADOX

There are several mechanisms that may contribute to the big data paradox arising in real-world studies, including RCTS:

1. Decreased Data Quality

As more patients enroll in a study, this might pose challenges like data overload and increased complexity in data integration from various sources, which could lead to lower data quality. With the former, studies can contain extensive amounts of data from various sources including patient records, lab results, and diverse measurements related to the intervention being studied – the sheer magnitude of information can be difficult to organize, analyze, and draw meaningful insights from without making too many generalized assumptions. In terms of complexity with data integration, studies may involve data from multiple sources and varied formats (i.e., clinical data, imaging, genetic information, patient-reported outcomes) which may lead to difficulties in combining these different datasets for comprehensive analysis. Multicenter RCTs that are conducted across diverse institutions and regions over long time periods may be particularly susceptible to this mechanism of the big data paradox (Msaouel, 2022).

2. Increased Patient Heterogeneity

As more patients are enrolled in a trial, there is increasing patient heterogeneity which may lead to increased bias because patient cohorts with elevated heterogeneity are more likely to harbor unmeasured characteristics that can increase irreducible error (Msaouel, 2022). For example, an unknown biomarker that was not measured but is assumed to be present may be included in a model and can influence parameter estimation – this would lead more error to be present in the model. Increased patient heterogeneity creates a setting where assumptions may need to be made to simplify data processing and analysis, sacrificing error minimization in the process.

3. Confidence Intervals Overlook Irreducible and Systematic Error

As the sample size of a study increases, standard error decreases. This means that for very large studies, total error is mostly due to irreducible and systematic error rather than standard error. Confidence intervals are traditionally a function of standard error. Thus, when they are calculated for large studies with high total error, they can be misleadingly narrow since they are primarily concerned with standard error.

4. Resource and Technological Demands

The integration and analysis of large amounts of study data can require advanced data analytics tools and expertise. Implementing the appropriate technologies, maintaining data security, and employing suitable analytical methods can pose technical challenges that need specialized knowledge and resources – without investing in these factors, data can be vulnerable to not being accurately analyzed, further exacerbating issues with patient heterogeneity and data quality.

STRATEGIES TO ATTENUATE THE BIG DATA PARADOX

It is essential to mitigate the impact of the big data paradox in research studies to have more accurate results. Some strategies that can be employed include:

1. Placing more **emphasis on data quality** rather than only focusing on data quantity. This can be done by establishing robust data collection methods, implementing stringent data cleaning processes, and monitoring for data accuracy, completeness, and consistency. Data quality can also be improved by *being more explicit* about the data that was used in analysis when reporting a study, including steps related to data selection, data processing, curation, and analysis (Msaouel, 2022).
2. Considering **data reduction and selection** by focusing on key variables or subsets of data that are most relevant to the research question. This would reduce the volume of data while maintaining the quality of the information that is being analyzed.
3. Anticipating and measuring the sources of **potential irreducible error** prior and during data collection which can be used to improve statistical models (Msaouel, 2022).
4. Employing **multilevel/hierarchical modeling** techniques to account for patient heterogeneity. Such models provide a framework to investigate for signals in a large data set while minimizing noise from information across multiple levels.
5. Implementing **advanced analytical techniques** that could better handle the diversity and complexity of large datasets – such as machine learning and AI algorithms.
6. Implementing **data visualization techniques** (i.e., graphs, charts, dashboards) to present complex data in a more accessible and understandable format and in identifying patterns and relationships within a dataset.

7. Replacing traditional **calculations of confidence intervals** with error intervals that account for both standard error and systematic error.
8. Fostering **interdisciplinary collaboration** between statisticians, data scientists, clinicians, and ethicists to develop comprehensive strategies that address characteristics of a study that could contribute to the big data paradox in research studies.

CONCLUSION

The big data paradox is becoming increasingly pernicious as contemporary advances allow for the integration of large scale data and clinical trials aim to enroll more patients to increase the relevance of their results for clinical practice. There are several mechanisms that lead to this phenomenon, including decreased data quality, increased patient heterogeneity, and the shortcomings of confidence intervals to account for total error. Mitigating the big data paradox is essential to elucidating more accurate results. Strategies to do this include emphasizing enhanced data quality, employing more adaptable statistical modeling to accommodate and address the greater diversity within larger patient groups, utilizing data visualization techniques, encompassing both systematic and standard errors in confidence intervals, and fostering interdisciplinary collaboration.

REFERENCES:

- Gandhi, L., Rodríguez-Abreu, D., Gadgeel, S., Esteban, E., Felip, E., De Angelis, F., Domine, M., Clingan, P., Hochmair, M. J., Powell, S. F., Cheng, S. Y.-S., Bischoff, H. G., Peled, N., Grossi, F., Jennens, R. R., Reck, M., Hui, R., Garon, E. B., Boyer, M., ... Garassino, M. C. (2018). Pembrolizumab plus Chemotherapy in Metastatic Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, *378*(22), 2078–2092. <https://doi.org/10.1056/NEJMoa1801005>
- Meng, X. L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. <https://doi.org/10.1214/18-AOAS1161SF>, *12*(2), 685–726. <https://doi.org/10.1214/18-AOAS1161SF>
- Msaouel, P. (2022). The Big Data Paradox in Clinical Practice. *Cancer Investigation*, *40*(7), 567–576. <https://doi.org/10.1080/07357907.2022.2084621>
- Reck, M., Rodríguez-Abreu, D., Robinson, A. G., Hui, R., Csőszi, T., Fülöp, A., Gottfried, M., Peled, N., Tafreshi, A., Cuffe, S., O’Brien, M., Rao, S., Hotta, K., Leiby, M. A., Lubiniecki, G. M., Shentu, Y., Rangwala, R., & Brahmer, J. R. (2016). Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, *375*(19), 1823–1833. <https://doi.org/10.1056/NEJMoa1606774>
- Senn, S. (2022). Empirical studies of balance do not justify a requirement for 1,000 patients per trial. *Journal of Clinical Epidemiology*, *148*, 184–188. <https://doi.org/10.1016/j.jclinepi.2022.02.010>

Submitted 11/7/2023

Section IV. Advanced Research Methods and Statistics

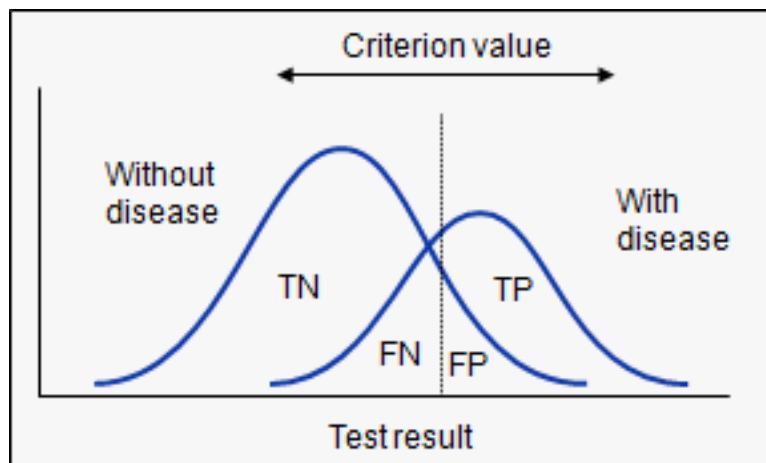
IV.1 Receiver Operating Curve Basics (Karim Farrag and Julia Lake)

Among the armory of statistical tools available to the evidence-based clinician, the **Receiver Operating Characteristic (ROC)** curve, and analysis of the Area Under the Curve (AUC), present an invaluable tool for measuring the how effective our tests are at distinguishing signal from noise.

Born out of the signal detection theory in World War II, the term “receiver operating characteristic” originates from the ability of radar operators to distinguish meaningful blips on the radar (enemy/allied vessels) from non-meaningful noise (birds, nature, etc...) [1]. Namely, the receiver operating characteristic served as a measure of those individuals’ ability to distinguish signal from noise, and since the 1970’s has been repurposed in the medical field to assess the efficacy of our tests. More or less, analysis of the area under the ROC curve seeks to address the important clinical question: how well does a given test distinguish disease from non-disease? To understand how this analysis is completed, we must first examine the different statistical components that go into creating an ROC curve, and how they are analyzed to produce a clinically useful AUC measurement.

At its core, the Receiver Operating Characteristic (ROC) curve is created for a given binary test by plotting the test’s true positive rate (TPR = test sensitivity) on the y-axis vs its False Positive rate (FPR: 1 – test specificity) on the x-axis. The corresponding curve that is plotted is known as the ROC curve, and the calculated area under that curve provides a number from 0 (a perfectly inaccurate test) to 1 (a perfectly accurate test) which tells us how effectively the test in question discriminates between the null hypothesis and its counterpart, or in the clinical realm, between disease and the lack thereof.

To understand this concept further, we must first consider two hypothetical populations, one with the disease and one without. Statistically, each population will present with (often overlapping) bell shaped distributions of symptoms or detectable signals (see figure A below). Our test in question (on which we will be applying ROC analysis), represents the dotted line drawn, which is our best attempt to distinguish between these two overlapping distributions of “disease” and “no disease”.



Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

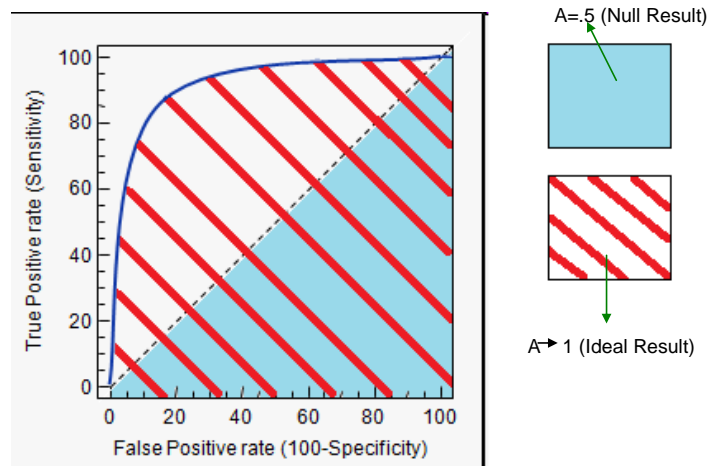
Figure A

The ROC curve analysis is the tool we use to determine how well this test actually distinguishes between the two distributions outlined (how well does it discriminate between disease and no disease). Two intrinsic values relevant here are sensitivity and specificity. For any test, we can determine how low the threshold signal needs to be to register as a positive hit (how sensitive our test is). This in turn comes at a cost of specificity: the wider we cast our net, the more likely we are to pick up false positives (and thus be less specific). The balance between how sensitive we want our test to be, and the amount of false positive rate we will tolerate is determined by many factors, including the nature/risks of the disease for which are testing (such as a cold vs cancer) and the technological limits of the test itself.

By plotting our test's True Positive Rate (Sensitivity) at the various accepted False Positive Rates ($1 - \text{Specificity}$), we create an ROC curve which provides us useful clinical information. Ultimately, the shape of said curve and the area under it (AUC) elucidates the effectiveness of our test. A perfect test has 100% sensitivity at all accepted false positive rates, which is to say that even if the test were set to have no false positives (i.e., be 100% specific) it would be 100% sensitive. The area under this curve would be 100% of the space in question, (i.e., 1). On the other had, if our false positive rate is equal to our true positive rate, then our binary test is only accurate 50% which is in fact no more accurate than chance. Thus, a corresponding AUC of 0.5 indicates that our test is not an accurate tool to distinguish between those with and without disease. Thinking in terms of the figure above, it means that the two distributions in question are perfectly such that no matter where we place our test's criteria, we are no better off than chance at discriminating between the two. This suggests that the tool we are using (such

ROC Curve Analysis

Area Under Curve (AUC): Measures strength of the test used in distinguishing between both distributions (i.e. how well does given feature distinguish disease vs the lack thereof ?)



<http://www.medcalc.org/manual/roc-curves.php>

as the symptom or marker we are testing) is not appropriate.

Figure B

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

AUC values ranging from 1 to 0 tell us how well our test is discriminating between these two populations. 0.5 indicates that our binary test is no more accurate than chance, and fails. In the realm of clinical medicine, a score of 1-.9 indicates an excellent test, .9-.8 a good test, .8-.7 a fair test, .7-.6 a poor test. Values lower than .4 indicate the test is consistently inaccurate, such that in theory, a perfectly inaccurate binary test (AUC of 0) would still clinically be perfectly discriminating between the diseased and non-diseased populations but is mislabeling the sick as healthy and the healthy as sick (which, in a binary system, is still helpful information).

Thus, in examining receiver operating characteristic analysis, we find a conceptually fascinating and tremendously useful tool which we may use to evaluate the efficacy of the tests which so often form the cornerstone of our clinical decision making.

Here's a brief review of key terms and concepts:

Definitions and Formulas

Sensitivity (True Positive Rate, TPR): the probability our test result is positive when the disease is present.

Specificity (True Negative Rate): the probability our test result is negative when the disease is not present.

False Positive Rate (FPR): the probability our test result is positive despite the absence of disease
 $= 1 - \text{Specificity}$.

Positive likelihood ratio: ratio between the probability of a positive test result in the *presence* of disease and the probability of a positive test result in the *absence* of the disease. = True positive rate / False positive rate = Sensitivity / (1-Specificity)

Negative likelihood ratio: ratio between the probability of a negative test result in the *presence* of disease and the probability of a negative test result in the *absence* of the disease, i.e., = False negative rate / True negative rate = (1-Sensitivity) / Specificity

Positive predictive value: probability that the disease is present when the test is positive (expressed as a percentage) = $a / (a+c)$

Negative predictive value: probability that the disease is not present when the test is negative (expressed as a percentage) = $d / (b+d)$

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

In tabular format:

Test	Disease				Total
	Present	n	Absent	n	
Positive	True Positive (TP)	<i>a</i>	False Positive (FP)	<i>c</i>	<i>a + c</i>
Negative	False Negative (FN)	<i>b</i>	True Negative (TN)	<i>d</i>	<i>b + d</i>
Total		<i>a + b</i>		<i>c + d</i>	

Table A

Where the following statistical formulas can be defined:

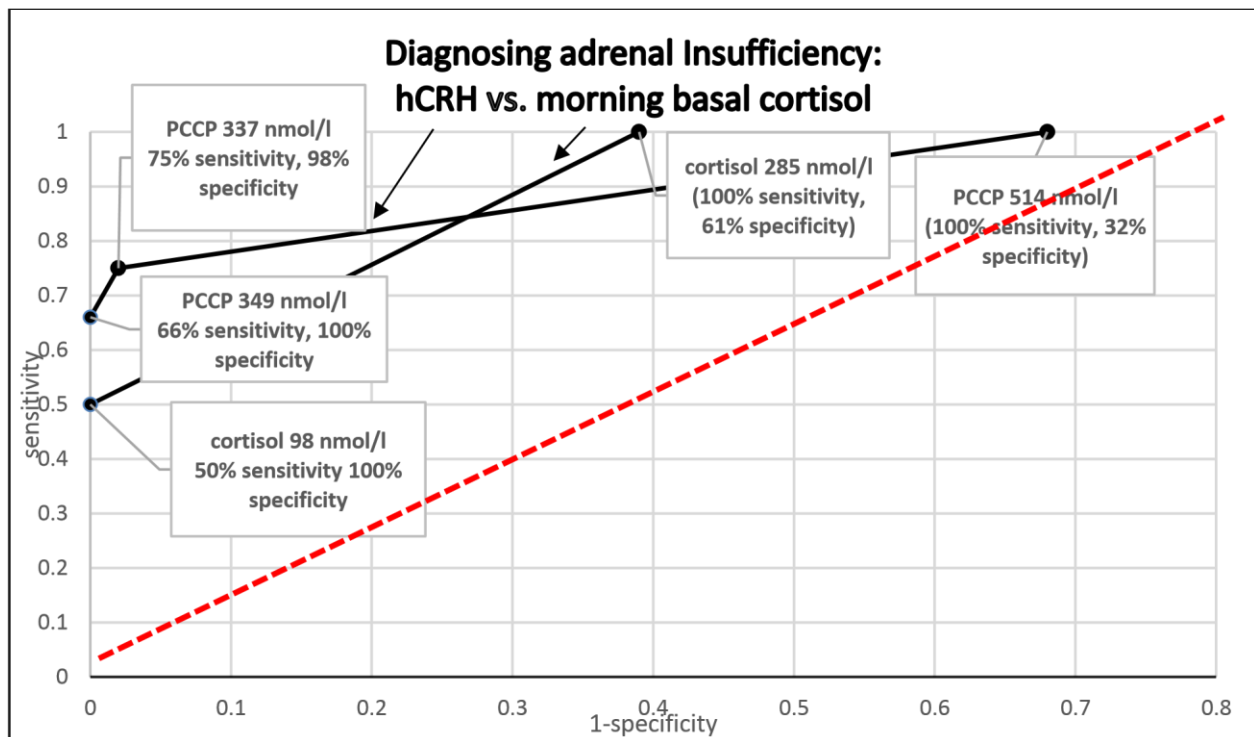
Sensitivity	$\frac{a}{a + b}$	Specificity	$\frac{d}{c + d}$
Positive Likelihood Ratio	Sensitivity/ 1 - Specificity	Negative Likelihood Ratio	1 - Sensitivity/ Specificity
Positive Predictive Value	$\frac{a}{a + c}$	Negative Predictive Value	$\frac{d}{b + d}$

Table B

B. How can I apply this to a clinical question? Consider the following

A ROC illustrates the **true positive rate** (sensitivity) on the y-axis and the **false positive rate** (1-specificity) on the x-axis at various threshold settings. A theoretical point in the left upper coordinate (0,1) of the ROC space reflects a test with **100% sensitivity** (zero false negatives) and **100% specificity** (zero false positives). Calculating both sensitivity and specificity against a gold standard at multiple points along this curve allows one to set an optimal value (cutoff) for a particular test.

For example, consider a ROC generated by comparing the sensitivity and false positive rate of a morning serum cortisol level for diagnosing adrenal insufficiency. Now, consider another ROC as compared to the ROC for the hCRH (human corticotropin releasing hormone) stimulation test. In this model, the gold standard insulin tolerance test is used to determine sensitivity and specificity.



The **tradeoff between sensitivity and specificity** is highlighted above. For instance, 98 nmol/l is the morning cortisol diagnostic level at which identifying adrenal insufficiency carries a sensitivity of 50%, but a specificity of 100%. When increased to 285 nmol/l, the sensitivity of this diagnostic threshold increases to 100%, but specificity decreases to 61%. Similarly, a peak cortisol cut point (PCCP) of 349 nmol/l obtained using hCRH is less sensitive but more specific than 514 nmol/l in diagnosing adrenal insufficiency. ⁽¹⁾ The most accurate diagnostic test is that which is closest to (0,1) or the point which is along the curve in the upper left-hand quadrant of the plot. The curve that describes an arc closest to that point (and has a large area under the ROC) is a “good” test, as opposed to one that is close to the line of unity, shown in the dashed line above.

Based upon these data, we can calculate positive and negative likelihood ratios to help our patients understand the likelihood they do or do not carry a diagnosis based upon their test result.

Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine

Diagnostic threshold	Positive likelihood ratio <u>Sensitivity</u> 1-specificity	Negative likelihood ratio <u>1-sensitivity</u> specificity
337 nmol/l	37.5	0.26
514 nmol/l	1.47	0.0003

In patient friendly words for instance, a person with a peak cortisol cut point of only 337 nmol/l, is much more likely to have adrenal insufficiency than someone with a peak cortisol cut point of 514 nmol/l based upon their positive likelihood ratios of nearly 38 and 1.5, respectively.

July 2017

References/Footnotes:

1. Schmidt, I. L., Lahner, H., Mann, K., & Petersenn, S. (2003). Diagnosis of Adrenal Insufficiency: Evaluation of the Corticotropin-Releasing Hormone Test and Basal Serum Cortisol in Comparison to the Insulin Tolerance Test in Patients with Hypothalamic-Pituitary-Adrenal Disease. *The Journal of Clinical Endocrinology & Metabolism*, 88(9), 4193-4198. doi:10.1210/jc.2002-021897.

Updated 5/1/2019

IV.2 Kaplan-Meier Curves- Comparing Event/Survival Between Experimental and Control Groups (Taeha Kim and Rachel Griffith)

What are Kaplan-Meier curves?

Kaplan-Meier curves are a graphical display of survival times which allows comparison of two or more groups. They consist of time on the X axis versus survival proportion on the Y axis. This allows for visualization and analysis of survival over time rather than at a single time point. For example, traditionally you may compare percent survival at one year for experimental versus control groups. With the Kaplan-Meier curve you can compare survival at all time points providing a more complete picture.

Can Kaplan-Meier curves only be used to analyze survival?

No. Kaplan-Meier curve can be used to analyze the amount of time it takes to reach any discrete event. For example, you could make a Kaplan-Meier curve displaying Time-to-MI for a new cardiovascular drug.

What is time-to-event?

Time to event starts either when the participant is recruited, or when treatment is given. It ends either when the event of interest occurs for the patient is no longer being followed. If a patient leaves the study without having the event of interest, this is called being censored.

Why are patients censored?

Patients can be censored for many reasons. Sometimes the study ends before the patient has the event of interest. Other times the patient chooses to leave the study prematurely and may be lost to follow up. Some patients will no longer be at risk for the event, and therefore it does not make sense to keep following them. For example, in a study of mortality in patients with class IV heart failure, someone who receives a heart transplant may be censored, because they are no longer have class IV heart failure.

Can the data from censored patients still be used?

Yes, this is one of the strengths of the Kaplan-Meier method. It allows incomplete data, that is, data from patients who do not reach the event to still be incorporated.

What are the assumptions made in constructing Kaplan-Meier curves?

1. Censored patients have the same survival as patients who reach the event of interest.
2. Survival probabilities do not depend on when a patient joins the study.
3. The time the event happens is the same as the time when it is detected.

When might the above assumptions not be true?

Sometimes patients are censored for a reason that makes them unlike the patients who reach the event of interest. If patients are censored simply because the study ends that may not indicate an inherent difference in the censored and uncensored populations. However sometimes patients drop out of a study for reasons that make them inherently different from those continue. Perhaps a treatment requires a great deal of motivation, or high health literacy. Patients who lack those characteristics may drop out, and they may also have worse survival secondary to the very characteristics that made them more likely to be censored.

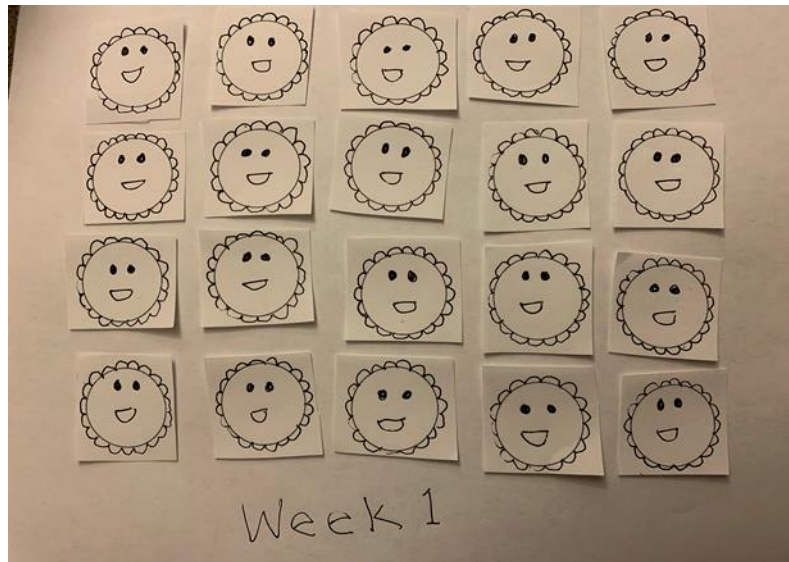
Sometimes survival probabilities do depend on when patients join the study. In some fast-moving fields, even a difference a few years makes mean that the later patients have access to more advanced treatments and have a better survival probability than those who came before them. Alternatively, there may be other outside factors affecting survival. You can imagine that a study that begin before the COVID-19 pandemic and continued into it may run into a problem with this assumption. Survival during the pandemic may be lower for reasons that have nothing to do with the study or the treatment.

Lastly the probability that an event is detected at the exact time it occurs depends on what that event is. Something like an MI or a hip fracture might be highly likely to be detected at the time it occurs. On the other hand, events such as ovarian cancer may take a very long time to be detected in the absence of screening. For these types of events, it is important that screening occurs frequently during the study period if Kaplan-Meier curves are to be used.

What's the whole idea behind the Kaplan-Meier method?

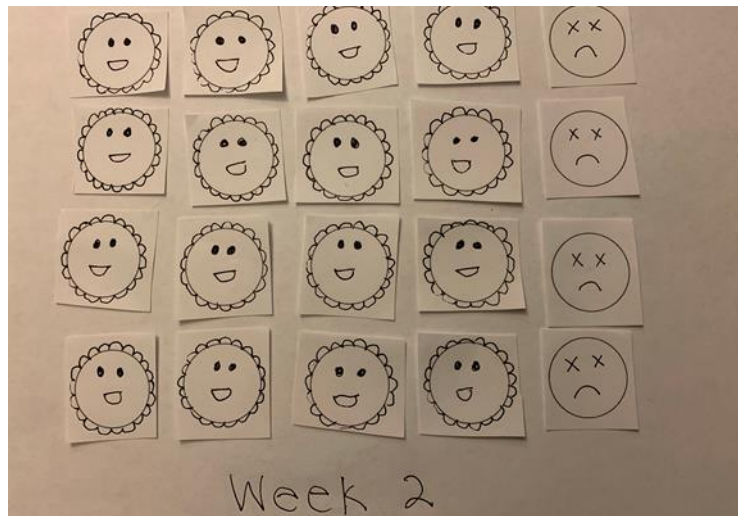
Let's say you want to do an experiment to see how long it takes for sunflowers to wither and die after they leave the store. You go to the store and buy 20 sunflowers. For the sake of simplicity, we'll assume you buy them all at the same time, but you could buy them at different times and in each case week 1 would be one week after you bought that particular sunflower. So, at the end of week 1 of your experiment you take a look at the sunflowers and they're all alive.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine



20/20 Surviving = 100%

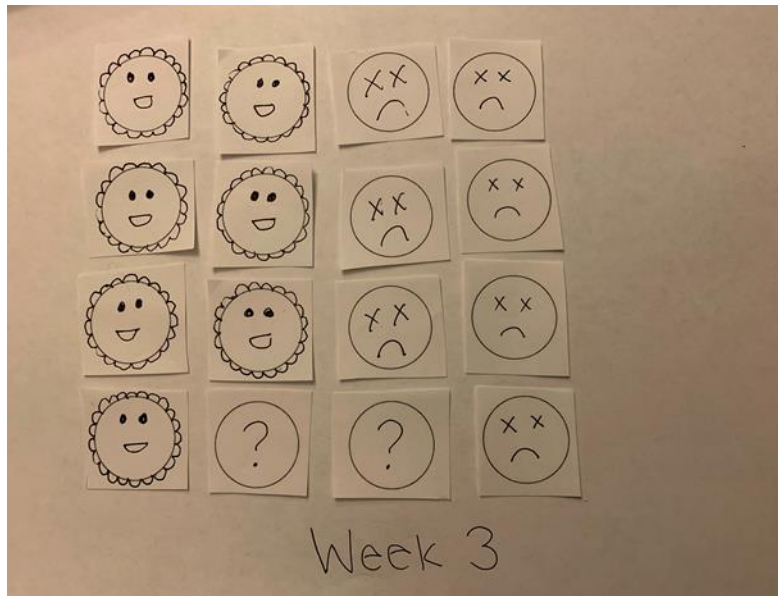
For the end of week 1, 20 plants were at risk of dying, and 20 survived, which means there was 100% survival. You wait a week and take a look again. This time, four have died, and sixteen are still alive.



16/20 Surviving – 80%

Between the beginning and the end of week 2, 20 plants were at risk of dying, and 16 survived, which means there was a 80% survival rate. On week 3, you again examine your sunflower plants. Seven are alive, and 7 are dead. Unfortunately, you find that 2 are missing! Your roommate explains that they gifted them to their parents. You know that they lived at least 2 weeks, but after they were given away you have no clue how much longer they lived. It could have been a day or a year. So instead of having an event, those two sunflowers have been censored.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine



7/14 Surviving = 50%

For week 3, we started with 16 plants... but we only have data for 14. So, of the 14 plants that were at risk of dying, 7 survived, or 50%. But we can't say that the plants have a 50% survival rate at the end of 3 weeks, because it wasn't a given that they would make it to the beginning of week 3. There was only an 80% chance of even being alive at the beginning of the week. So, to account for that, we multiply 80% by 50% or 0.8×0.5 . This is equal to 0.4, or 40%. So, the 3-week survival of your sunflower plants is 40%.

You may be tempted to take a shortcut and divide the 7 surviving plants by the 20 total starting plants. But that doesn't work, because it implies that our 2 missing plants are dead, and we just don't know whether that is true or not. Or maybe you would try to exclude them entirely and divide 7 by 18. But that would imply that those two plants didn't contribute any data... which isn't true because we know that they survived at least two weeks. By calculating survival during individual time intervals, and then multiplying them together, the Kaplan-Meier method allows us to incorporate incomplete data by assuming that once participants leave the study they will behave in the same way as those who are still in the study.

How do you actually do a full Kaplan-Meier analysis?

In evaluating a study that looks at an intervention and its effect on primary outcomes over time (e.g., death, first hospitalization, other markers of decline), there are 3 components to the statistical analysis that are commonly performed:

- 1) Construction of survival curve for experimental and control group
- 2) Calculation of the test statistic to determine statistical significance of difference between those two curves
- 3) Regression analysis to account for explanatory variables such as age and co-morbidities

Construction of survival curve: Kaplan-Meier Method

Kaplan-Meier Method is a way to construct survival curve as a function of time, $S(t)$, based on observational data. It looks at time intervals between events as they happen and takes a product of rate of survival for each interval. $S(k) = p_1 \times p_2 \times p_3 \times \dots \times p_k$, where $p_i = (r_i - d_i)/r_i$. r_i is the number alive at the beginning of period i and d_i the number of deaths within the period.

Comparing survival curves: log-rank test

Once the survival curves have been constructed, we need to compare them to determine whether there are statistically significant differences between them. We can do this by figuring out the test statistic (χ^2). The assumption is that there is no difference between the curves when you start (null hypothesis). Based on χ^2 , you can determine the P value and presence of statistically significant difference between the curves.

$$\chi^2 (\text{log rank}) = (O_c - E_c)^2/E_c + (O_e - E_e)^2/E_e$$

where O_e and O_c denote total number of observed events in experimental and control group, respectively, and E_e and E_c denote total number of expected events in experimental and control group, respectively.

Accounting for other explanatory variables: Cox’s proportional hazard method (Cox regression)

As you can imagine, survival is dependent not only on intervention being studied, but patient characteristics as well (e.g., age, BMI, co-morbidities). Cox’s proportional hazard method allows you to account for these other explanatory variables by making the ‘hazard’ the response variable and determine which of the explanatory variables being tested are actually relevant.

This model can be described as follows:

$$\ln h(t) = \ln h_0(t) + b_1x_1 + \dots + b_px_p$$

where $h(t)$ is the hazard function at time t , x_p denote explanatory variables and b_p denote coefficients that can be estimated from the observed data.

Since one cannot measure the instantaneous risk of death, we will have to use cumulative hazard function and perform regression w/ the cumulative hazard function:

$$H(t) = -\ln S(t)$$

where $S(t)$ is the cumulative survival function based on observation.

Example

Let’s suppose we ran a trial comparing treatment 1 vs treatment 2 and looked at the survival in patients receiving these interventions. The raw data is as follows

Patient #	Survival time (days)	outcome	treatment	Age
1	1	Died	2	67
2	3	Died	2	66
3	3	Unknown	2	75

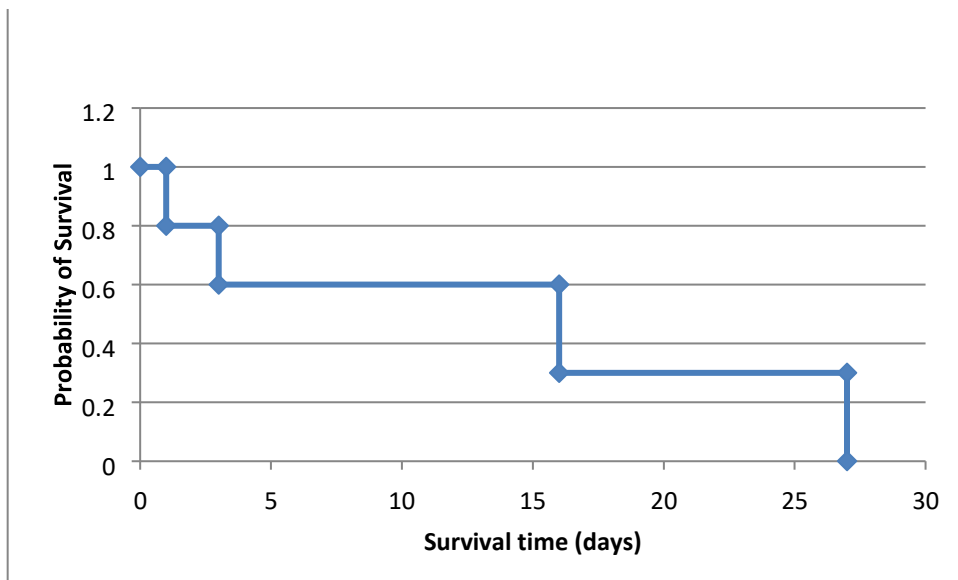
Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

4	7	Died	1	69
5	16	Died	2	81
6	25	Died	1	77
7	27	Died	2	66
8	39	Unknown	1	63
9	45	Died	1	75
10	77	Survived	1	72

We can calculate survival function, $S(t)$ for treatment 2 as follows

Patient #	Survival time (days)	# known to be alive (r_i)	Deaths (d_i)	Proportion surviving (p_i)	Cumulative surviving ($S(t)$)
	0				1
1	1	5	1	$(5-1)/5 = 0.8$	$1 \times 0.8 = 0.8$
2	3	4			
3	3+	4	1	$(4-1)/4 = 0.75$	$0.8 \times 0.75 = 0.6$
5	16	2	1		
7	27	1	1	$(1-1)/1 = 0$	$0.4 \times 0 = 0$

Now we are ready to plot the Kaplan-Meier survival curve.



Kaplan-Meier Survival Curve

You can repeat the same process to generate the Kaplan Meier survival curve for treatment 1 as well. Once you have those two curves, you can use the log-rank test to calculate the test statistic, which allows you to figure out whether there is a statistically significant difference between the curves. For this, you need to figure out expected deaths for each time interval. The assumption is that risk of death is same between the two groups (null hypothesis).

Survival time (days)	Treatment group	# know to be alive (r_i)	Deaths (d_i)	Risk of death (d_i/r_i)	# known to be alive from treatment group 2 (r_2)	Expected # of death in treatment group 2 (E_2)
0						
1	2	10	1	$1/10 = 0.1$	5	$5 \times 0.1 = 0.5$
3+	2	9				
3	2		1	$1/9 = 0.11$	4	$4 \times 0.11 = 0.44$
7	1	7	1	$1/7 = 0.14$	2	$2 \times 0.14 = 0.28$
16	2	6	1	$1/6 = 0.17$	2	$2 \times 0.17 = 0.34$
25	1	5	1	$1/5 = 0.2$	1	$1 \times 0.2 = 0.2$
27	2	4	1	$1/4 = 0.25$	1	$1 \times 0.25 = 0.25$
39+	1	3	0	$0/3 = 0$	0	$0 \times 0 = 0$
45	1	2	1	$1/2 = 0.5$	0	$0 \times 0 = 0$
77+	1	1	0	$0/1 = 0$	0	$0 \times 0 = 0$
						$E_2 = 2.01$

There were 7 deaths in the trial, and as E_2 is 2.01, it follows that $E_1 = 7 - E_2 = 4.99$.

Now we have all the O_1 , O_2 , E_1 and E_2 that we can plug in to figure out χ^2 . Using chi-square distribution table and one degree of freedom, you can figure out the P value.

The last step is performing Cox regression. Using the first table we created to generate K-M survival curve, we can generate cumulative hazard function as previously explained:

$$H(t) = -\ln S(t)$$

You will need to use a software to perform Cox regression, and that will give you a coefficient for each variable being tested, along with the P value and confidence interval. This allows you to determine which of the variables being tested are statistically significant.

What are some of the things to look for in Kaplan-Meier curves?

1. Note how large the steps are. A smaller number of larger steps means a smaller sample size, whereas a large number of smaller steps (producing the appearance of a smoother curve) demonstrates a larger sample, which is preferable.
2. Check to see how many subjects were censored— if many, then it might mean high attrition. As in any study this raises the risk of bias if those lost to follow up had different characteristics than those who continued.
3. Look for curves that show the number of participants at risk below each interval on the x-axis. There will be a minimum follow up time, where no participants have yet been censored, data from this time period and earlier is the most accurate. On the far right of the graph, the sample size may be very small, with only a few participants still being followed. These participants can have an outsized influence on the data. For example, maybe a drug has 50% survival at 9 years. At year 10 there is only 1 participant left, so survival for that interval will either be 100% or 0%. So the 10 year survival will be 50% if that person survives (0.5×1.0) but 0% if they die. That's a huge difference based on the data from only 1 person.

References:

1. Bland, J M., and D. G Altman. "Statistics Notes: Survival Probabilities (the Kaplan-Meier Method)." *BMJ*, vol. 317, no. 7172, 1998, pp. 1572–1580., <https://doi.org/10.1136/bmj.317.7172.1572>.
2. Kishore, Jugal, et al. "Understanding Survival Analysis: Kaplan-Meier Estimate." *International Journal of Ayurveda Research*, vol. 1, no. 4, 2010, p. 274., <https://doi.org/10.4103/0974-7788.76794>.
3. Rich, Jason T., et al. "A Practical Guide to Understanding Kaplan-Meier Curves." *Otolaryngology–Head and Neck Surgery*, vol. 143, no. 3, 2010, pp. 331–336., <https://doi.org/10.1016/j.otohns.2010.05.007>.
4. Stel, Vianda S., et al. "Survival Analysis I: The Kaplan-Meier Method." *Nephron Clinical Practice*, vol. 119, no. 1, 2011, pp. c83–c88., <https://doi.org/10.1159/000324758>.

Submitted 11/22/16 (T. Kim) and updated 12/20/2021 (R. Griffith)

IV.3 The Cox Proportional Hazards Model (Art Kehas)

The Cox proportional hazards model allows you to model the simultaneous effect of several factors, or *covariates*, on an outcome at a particular point in time. For instance, if you wanted to look at the effect of smoking, sex, and age on incidence of myocardial infarction or perhaps survival over a time period, you would use a Cox proportional hazard model to do so. This is similar to a Kaplan-Meier curve but different in that a Kaplan-Meier curve only models an outcome with respect to one variable.

Mathematically, the Cox model is expressed as a *hazard function*. To better understand the statistics involved, please access the following citations. Simply put, the output of the hazard function is a *hazard ratio* (HR). Thus, like most ratios, we can define three general scenarios: (1) $HR = 1$; (2) $HR < 1$; and (3) $HR > 1$. A $HR = 1$ denotes no effect of the covariates on the outcome. A $HR < 1$ shows a reduction in the probability (or hazard) of the outcome occurring. Finally, a $HR > 1$ denotes an increase in the probability (or hazard) of the outcome occurring. As always, hazard ratios will include a confidence interval. If the confidence interval crosses 1, then the result is not significant.

https://www.statsdirect.com/help/survival_analysis/cox_regression.htm

<http://www.sthda.com/english/wiki/cox-proportional-hazards-model>

IV.4 Aggregation of Data in Meta-Analyses and How to Assess for Robustness of the Results (Chelsea Gaviola, GSM4)

Meta-analyses combine multiple studies and, with the larger sample size, can provide more power than individual studies alone. It can be attractive to look at meta-analyses as they summarize the results of more studies, provide a big picture overview, and may settle conflicting results. However, as with any methodology, there are important limitations to keep in mind.

In this section, we hope to answer the following questions: How do meta-analyses collate and analyze data from multiple studies? What assumptions are made in doing so? How do you know if the results are applicable and accurate?

It would be helpful to review the chapters on [“Systematic reviews and meta-analysis,”](#) [“Assessing risk of bias of randomized controlled trials in systematic reviews and meta-analyses,”](#) [“Heterogeneity,”](#) and [“Forest Plots”](#) prior to this section.

How do meta-analyses analyze data from multiple studies and what assumptions are made?

There are two stages to meta-analyses. In the first stage, a summary statistic is calculated for each study, to describe the observed intervention effect (e.g., relative risk) in the same way for every study.

In the second stage, a combined intervention effect estimate is calculated. There are two major models used to pool data from individual studies: the **fixed-effects model** and the **random-effects model**.

The **fixed-effects model** is based on the following formula:

$$\text{generic inverse – variance weighted average} = \frac{\sum Y_i \left(\frac{1}{SE_i^2} \right)}{\sum \left(\frac{1}{SE_i^2} \right)}$$

where Y_i is the intervention effect estimated in the i^{th} study and SE_i is the standard error of that estimation, and the summation is across all studies. This model assumes that **the intervention effect is the same** across all studies and **tries to find the true underlying effect**. Larger studies have smaller standard errors and contribute more weight than smaller studies, which have larger standard errors. This model implies that the observed differences among individual study results are solely due to chance.

The **random-effects model** assumes that the **intervention effect is not the same** across studies. It assumes that the individual studies are estimating different, but related, intervention effects, that follow some distribution (usually a normal distribution). The final value represents the **average effect across all studies**. In this model, each study is weighted equally. This model implies that the observed differences among individual study results are due to both chance and some true variation in the intervention effects. The random-effects calculation relies on the standard errors of the study-specific estimates (SE_i), which are adjusted to incorporate a measure of the extent of heterogeneity.

If there is no heterogeneity among studies, both models will have identical results. If there is heterogeneity, the confidence interval will be wider in the random-effects model.

Fixed-effects model	Random-effects model
Intervention effect is the same. Tries to find the true underlying effect.	Intervention effect is different but related and follows a distribution pattern. Tries to find the average effect.
Studies with larger sample size (smaller standard error) are given more weight.	Studies are weighted equally. However, relative to the fixed-effects model, smaller studies have more weight.
Implies differences between individual study results are due to chance.	Implies differences between individual study results are due to chance and true variation in the intervention effects.
Smaller confidence interval.	Wider confidence interval.

Because the fixed-effects model operates under the assumption that there is some true underlying intervention effect, it ignores heterogeneity among studies. It is important for the authors of fixed-effects meta-analysis to investigate heterogeneity and include a discussion in the results (see how by reviewing Chapter 14 by Richie Huynh).

Smaller studies are weighted relatively more in random-effects than in the fixed-effects model. This can pose a problem if the results of smaller studies are different from the results of larger ones. If this is the case and there is funnel plot asymmetry suggesting a relationship between intervention effect and study size, then the random-effects model will be skewed towards the findings of smaller studies. (For a review on funnel plots, see [the chapter by Chris Lindholm](#)).

The best approach would be to present both models in a meta-analysis, with a funnel plot and sensitivity analysis/fragility index to show the strength of the results.

How reliable are the results? Looking for sensitivity analysis and the fragility index

A **sensitivity analysis** is helpful to determine the strength of the aggregated results. One common way to perform this analysis is to conduct a repeat meta-analysis without inclusion of the studies that you suspect may bias the results. For example, if the original meta-analysis included published and unpublished studies, you could re-conduct the analysis without the unpublished studies and see if the results are still consistent with the first analysis. If the results change, then the original meta-analysis results are less credible. The next time you are reviewing a meta-analysis, see if the authors conducted a sensitivity analysis and see if the results remain consistent among the different subgroups.

A newer method to evaluate the strength of a meta-analysis's conclusions is to calculate the **fragility index**, which is a concept discussed in a 2019 paper by Atal et al. The fragility index is a known concept for RCTs and is defined as a minimum number of non-events that would need to be changed to events in one arm to switch the result to statistically insignificant. The authors adapted this definition for meta-analyses and determined its fragility index to be the minimum number of patients from one or more trials included in the meta-analysis for which a modification in the event status (i.e., changing an event to a nonevent, or a nonevent to an event) would change the statistical significance of the pooled intervention effect to statistically insignificant. The authors in this study analyzed 906 meta-analyses (400 had statistically significant and 506 had insignificant results). For the statistically significant meta-analyses, the median fragility index was only 12. Overall, they found that the statistical significance of 33% of all meta-analyses depended on the status of 5 or fewer patients from one of more specific trials. If authors of meta-analyses begin calculating and including a fragility index, it could be another marker of the strength of the results, and we could then place more trust in meta-analyses with higher fragility indexes.

In summary, there are a lot of statistical gymnastics and assumptions that go into meta-analyses. It is appealing that they aggregate the results of multiple studies, but there are pitfalls if the studies and study populations are disparate. If there is a large, well-designed RCT with a study population that better represents your patient, the RCT results are likely more reliable. If such an RCT does not exist, remember to analyze the meta-analysis for sources of bias (as described in [the chapter by Chris Lindholm](#)), and look for sensitivity and fragility indexes to assess for the strength of its conclusions.

References:

1. Atal I, Porcher R, Boutron I, Ravaud P. The statistical significance of meta-analyses is frequently fragile: definition of a fragility index for meta-analyses. *J Clin Epidemiol*. 2019 Jul;111:32-40. doi: 10.1016/j.jclinepi.2019.03.012. Epub 2019 Mar 30. PMID: 30940600
2. Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
3. De Luca G, Suryapranata H, Stone GW, Antoniucci D, Neumann FJ, Chiariello M. Adjunctive mechanical devices to prevent distal embolization in patients undergoing mechanical revascularization for acute myocardial infarction: a meta-analysis of randomized trials. *Am Heart J*. 2007 Mar;153(3):343-53. PMID: 17307410
4. Walker E, Hernandez AV, Kattan MW. Meta-analysis: Its strengths and limitations. *Cleve Clin J Med*. 2008 Jun;75(6):431-9. Review. PMID: 18595551

Submitted Feb. 2020

IV.5 Heterogeneity (Richie Huynh)

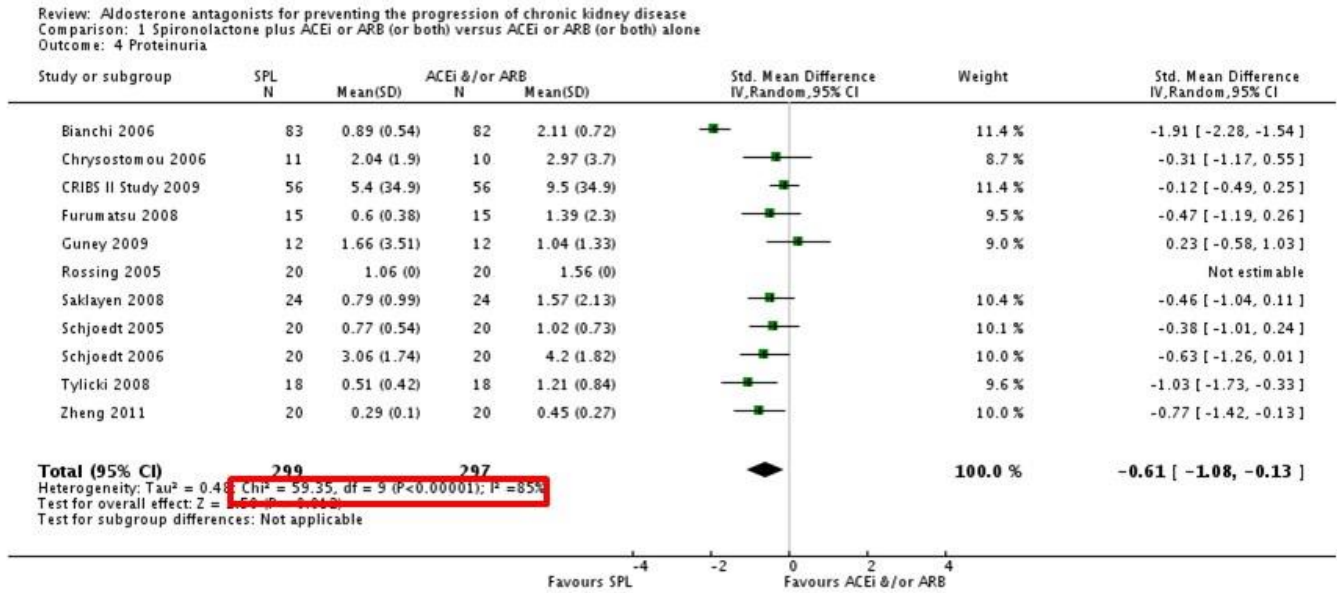
A systematic review summarizes the clinical literature from a systematic search, critical appraisal, and synthesis of the known worldwide literature on a specific issue. As such, when systematic reviews analyze and summarize this data, this is called a meta-analysis. To that end, an ideal meta-analysis would have homogeneity, meaning that the multiple studies being analyzed are appropriately similar and, thus, would be valid and helpful for comparison and aggregate data analyses.

The Cochrane Q and I^2 statistic together assess heterogeneity, or *true variations* among studies likely *not* due to chance. Thus, if there *is* heterogeneity, it'd be like comparing platypuses to direwolves, or apples to oranges, and thus not very helpful or valid for drawing conclusions.

- Classically, heterogeneity was measured **with Cochran's Q , also called $\chi^2(Q)$** , a weighted sum of squared differences among individual studies. The power of Q is dependent on the number of studies (N) and can be overpowered with a large N or conversely underpowered with a small N.
- To minimize said noted variations in power, heterogeneity is now standardly also assessed with the **I^2 statistic** to test for heterogeneity (*NOT to be confused with χ^2 or Q!*).
 - **$I^2 = (Q - df) / Q \times 100$** , where Q is Cochran's χ^2 and df is degrees of freedom.
 - I^2 is denoted as a %. If $df > Q$, I^2 is often denoted as 0% (rather than a negative I^2).
 - I^2 of $>50\%$ is generally considered **high** heterogeneity; some studies define their I^2 , such as an I^2 of 25%, 50%, and 75% as low, medium, and high heterogeneity, respectively.
- **When we assess heterogeneity, we look at Q first, then its P-value, and then the I^2 statistic:**
 - First, identify Q, or χ^2 - or chi-square (all the same thing). That's step 1.
 - Ask if the P value for Q is significant. The level of significance for Q is often set at 0.10 because of the low power of the test to detect heterogeneity. So, ask is $P > 0.10$? A *high* P-value is "good" and *ideal* as that suggests that heterogeneity is *insignificant*. In this case, we want to see that heterogeneity is *not* statistically meaningful, so we want a P value that is >0.10 , which is often "backwards" from what we're used to seeing and looking for with small P values <0.05 . Thus, in the case of heterogeneity, if P is <0.10 , then that suggests that the heterogeneity present *is* meaningful, which weakens the applicability of the meta-analysis because that indicates the studies are very, very different (and thus comparing orangutans to capybaras, or apples to oranges!).
 - Next, look at the I^2 to see how much heterogeneity there is. Is $I^2 > 50\%$? Is it super-high? The higher the I^2 , the more heterogeneity there is.
 - 🔗 If there *is* heterogeneity, a good systematic review would at least attempt to explain or speculate why that exists and how they may or may not account for it with their data analysis, ie, subgroup analysis, etc.
- Let's apply this to an example, to assess only for heterogeneity:

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Let's walk through this together, focusing on the #'s in the **highlighted red box**:



Cochrane Database of Systematic Reviews

29 APR 2014 DOI: 10.1002/14651858.CD007004.pub3

<http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD007004.pub3/full#CD007004-fig-00104>

1. What is Q? Here, Q is 59.35. This is the same as “chi².”
2. Is Q significant? The P-value is <0.00001. Yikes... remember we “want” a high P-value >0.10, but here this suggests that there *is* heterogeneity and that it is very significant.
3. What is the I²? Looks like I² is 85%, consistent with high heterogeneity (and >50%).
-What if the study did *not* list I²? Well, remember that $I^2 = (Q - df) / Q \times 100$.

So, we can always calculate I² manually.

Just for kicks, let's calculate that now: $[(59.35 - 9) / 59.35] \times 100 = 85\%$

➤ Now you can apply this to Forest Plots in systematic reviews and meta-analyses! Hooray!

August 2017

IV.6 Forest Plots (Richie Huynh)

Q: What is a Forest Plot?

A: In short, it's a graphical summary of a meta-analysis in systematic reviews. It's a unique graphical representation of *multiple* studies regarding an intervention effect and each studies' confidence intervals (CI). These were developed initially specifically for medical literature and is also called **blobbograms** but only by those who call X-rays roentgenograms instead of X-rays.

Q: So why should I care about blobbograms- uh, I mean Forest Plots?

A: Because it's a quick way to visualize the summarized results of systematic reviews, and you'll also see it a lot during Journal Clubs, and the rest of your career. It'll save you a lot of time if you actually know what you're looking at. Because it's a comparison of individual studies, this needs to be presented in a standardized way, and often times this will be relative risk (RR) or odds ratios (OR).

Q: OK, fine - so I have to return, like, a bunch of pages – what do I really need to know to do this?

A: Sure, when you look at it, look for 4 things:

1. The horizontal lines with squares – each line represents an individual study and its standard mean difference (95% CI), with the center square sizes indicating the weight of that specific study to the entire meta-analysis.
2. The solid **vertical line** represents where the intervention had no effect, which is an OR/RR of 1. Thus, left of that solid line supports the intervention and the right side favors the control.
3. The **Black Diamond** at the bottom (think skiing!) represents the average effect of all the combined studies in the meta-analysis. This usually is centered on a **dotted vertical line**.
4. Look to see if there is **significant heterogeneity** for the studies. What is heterogeneity, and when is it significant? If you've forgotten or want a refresher, review chapter 8 above!

Q: OK, I see those three things. What do they mean? How do I interpret them? Quickly!

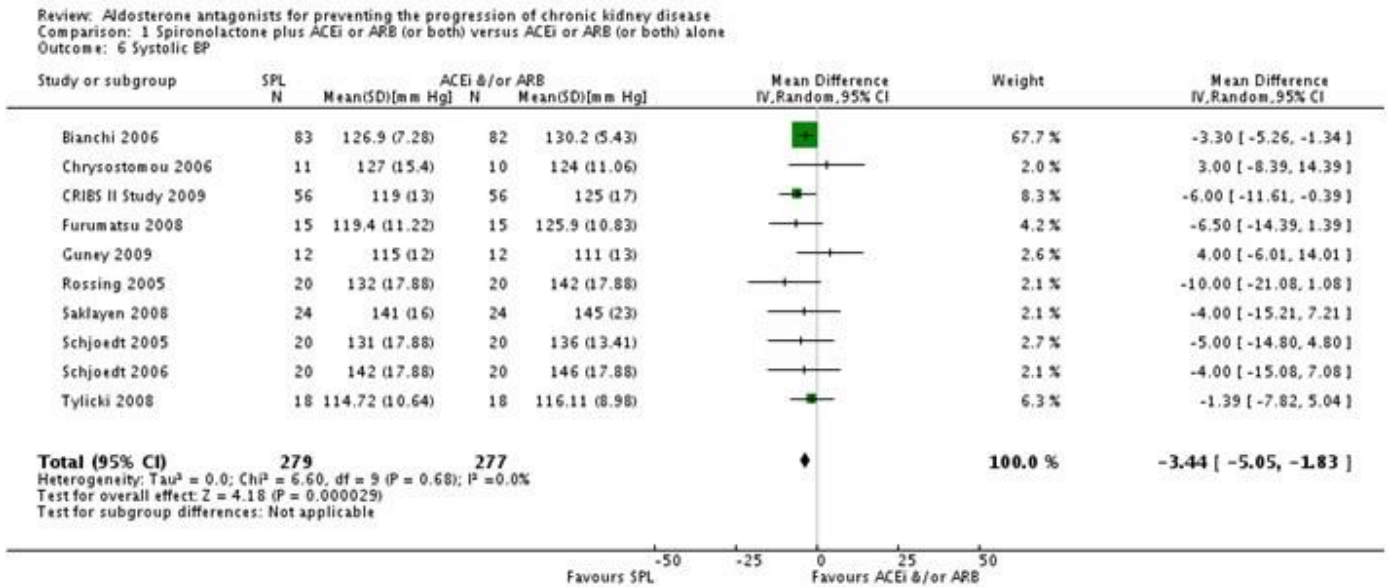
A: Well, you're halfway there! Ask yourself the following questions in order to determine the direction and magnitude of the intervention vs control (positive vs negative, small vs large).

1. Where do most of the individual studies line up? Look for the **dotted vertical line** that represents the average of all the studies, which **should line up with the Black Diamond**. *Is it to the left or right of the solid vertical line, so does it support the intervention or favor the control?*
2. *When you look at the horizontal lines representing individual studies, do their CI's cross and overlap the **solid vertical line**?* When the CI's overlap with the no effect line (OR/RR of 1), that means they're not statistically significant.

3. *Is there heterogeneity?* Look at the **Q** and its **P-value** and then the **I²** to determine whether there is significant heterogeneity (Chapter 8 above). You can also *qualitatively* visualize this with an “eyeball test” by seeing if the CI’s of the individual studies line up well or not, and oftentimes the dotted vertical line will cross all of the horizontal lines’ CI’s if the meta-analysis does *not* have significant heterogeneity.

Q: Can we practice this? Because I don’t really get it unless we apply it to real life, you know?

A: I hear ya, I’m the same way. OK, let’s look at an example:



Cochrane Database of Systematic Reviews
 29 APR 2014 DOI: 10.1002/14651858.CD007004.pub3
<http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD007004.pub3/full#CD007004-fiq-00106>

Let’s walk through this together, just like reading an EKG or a CXR.

1. First, I see the **horizontal lines** representing individual studies and their CI’s. I see the **black diamond** which should line up with a **dotted vertical line** that is actually missing here, but I can envision it. The imaginary dotted vertical line / black diamond is just to the *left* of the **solid vertical line** representing no effect between the intervention and control (OR/RR of 1). The intervention in this example is utilizing spironolactone as an addition to an ACEi or ARB or both, compared to a standard of just an ACE or ARB or both for lowering systolic blood pressure (SBP). Since the imaginary dotted vertical line and black diamond is to the left, this means the meta-analysis supports the intervention of adding spironolactone for lowering SBP.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

2. Additionally, I can see that the **largest square** is the Bianchi 2006 study, meaning that study contributed the most, or had the greatest weight, to the entire meta-analysis. If we look at the CI's, we can also see that the CRIBS II 2009 study in addition to the Bianchi 2006 study are the only two studies that are statistically significant because their **CI's do not cross the solid vertical line**; the other studies' CI's all cross this no-effect line.
3. Next, I look for heterogeneity. Remember, this is the **Q (chi²) and its P-value and the I²**.

Here, Q is 6.60, and P is 0.68, which is >0.10, meaning that heterogeneity is *not* significant. Next, I look to the I² to confirm this, and it is indeed 0%, meaning no statistical heterogeneity at all. Remember, we can manually calculate this, with **I² = (Q-dF) / Q x 100**. So, here it is: [(6.60 - 9) / 6.60] x 100 = -36.6%, but, remember, we should just denote it simply as I² = 0% when dF > Q, rather than a negative I².

Good job, you did it! You've interpreted the forest plot of this meta-analysis across ten whole studies all at once!

Q: Yes, I did, didn't I? ...OK, thanks and see you later at noon conference!

A: No problem. Yeah, see you then! Have a nice day!

August 2017

IV.7 Regression analysis – a. Introduction to Linear Regression Analysis (Jiyong Lee)

Linear regression attempts to predict the relationship between two variables by fitting a linear equation based on observed data. In other words, it describes the relationship between one dependent variable and explanatory variables.

Using regression starts by asking three questions.

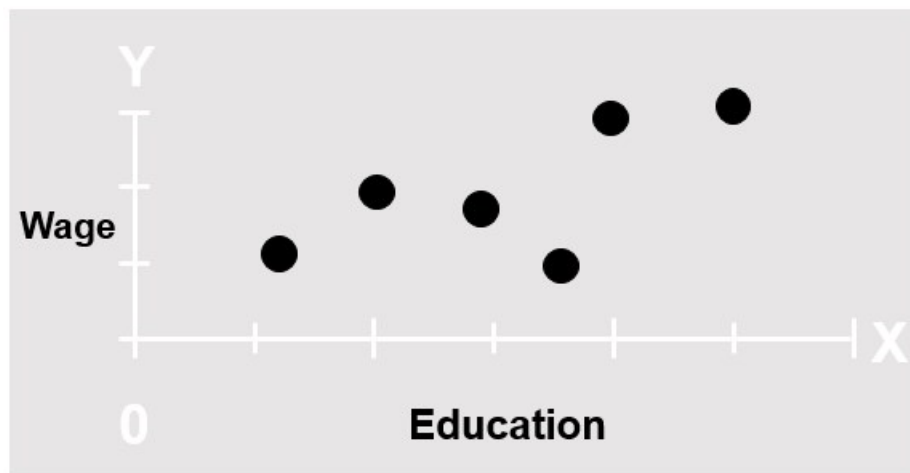
1. What is the sample regression line?
2. What is prediction of value of the dependent variable in relation to the independent variable?
3. Is there a statistically significant relationship between the independent and the dependent variables?

A Case –

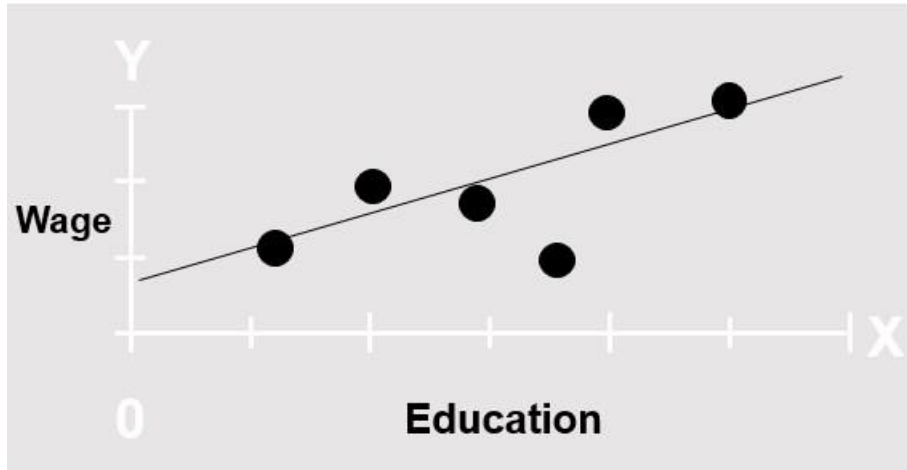
Let's learn it by doing a case. Our question will be "Is higher education associated with increased salary?"

We will get our data by surveying 100 people, assuming that the surveyed population will represent the whole population accurately. The population regression is the line of best fit using everyone in the population. However, it is impossible to gather data from everyone on the earth. So we obtain a random sample. If the sample is large enough and the observations are randomly selected, then the sample regression line should be a good predictor of the population regression model.

Then, you create a scatter graph.



The dependent variable (wage) is on the y-axis while the independent variable (years of education) is on the x-axis. In the next step, you draw a line of best fit.



The equation of the line is $y=mX+b$ (you remember that, right?). m represents the slope of the line and b presents the point where the line cuts the y -axis.

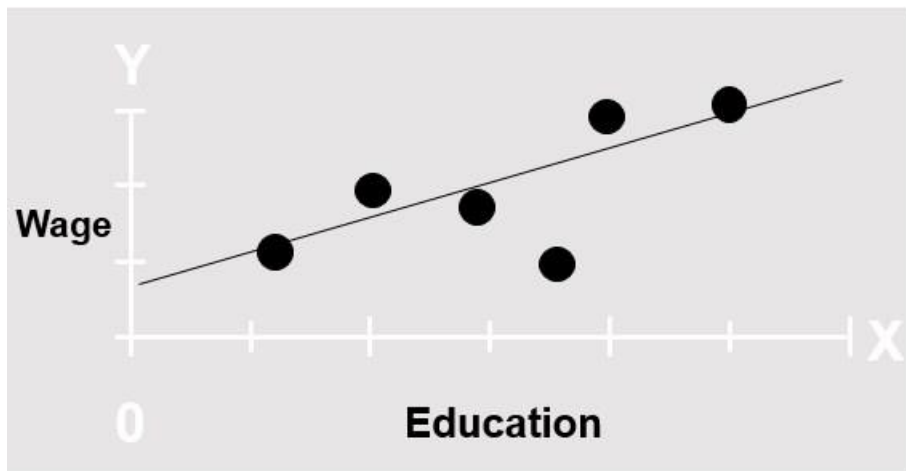
In statistics, the equation is presented as $Y=B_0 + B_1X$.

When B_1 (or m) is positive, it represents positive relationship.

When B_1 is negative, it represents negative relationship.

When B_1 is Zero, it implies no relationship.

Many statistical analysis tools including Excel and Stata can predict the equation for us. The most common method for fitting a regression line is the **method of least-squares**. This method calculates the line that fits best for the observational data by minimizing the sum of the squares of the vertical deviations from each data point to the regression line (if a point lies on the fitted line, then vertical deviation is zero).



Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

In this graph,

$$\text{Wages} = Y = B_0 + B_1X$$

$$= 7 + 1x \text{ \# of education years}$$

Here, the relationship showed that wage is expected to increase by \$1 per hour for every 1 additional year of education. \$7 is minimal wage. In brief summary, the regression line is the “line of best fit”. B_1 is the slope of the line. B_0 is the value of Y when X is equal to zero. The estimated regression can be used to make predictions for Y given X .

#Residuals

When the salary of a person is calculated with the equation provided above, it will not produce an exact wage. The discrepancy between prediction and actual wage is called residuals. In equation, it can be described as $Y = B_0 + B_1X + E$. E represents other factors contributing to wages including experience, job market, location, supply and demand, etc.

#Sum of squared error

SSE is the sum of all the residuals squared. The SSE is the measurement to determine how well the estimated line fits the observational data. Small errors represent better (more accurate) line.

IV.8 Regression Analysis- b. Understanding and Using Logistic Regression (Muhammad Khan GSM4)

What happens when quantitative response variables cannot be simply expressed as linear function of any explanatory variable? It is sometimes possible to reveal a linear relationship by using a simple mathematical transformation, such as computing the logarithm of one or both quantitative variables. In this chapter we will be reviewing some of the basics of **Logistic regression**.

Logistic regression is used where the response or dependent variable is categorical. It can be a *Yes* or *No* questions, such as , “How does the probability of getting cancer (yes vs. no) change for every pack of cigarettes smoked per day?” or “Do body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack (yes vs. no)?”

Moreover, the dependent variable should be dichotomous in nature and there should be no outliers in the data. The idea of Logistic Regression is to find a relationship between features and probability of particular outcome. Like all regression analyses, the logistic regression is a predictive analysis. In summary, logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

The **logistic regression** can be expressed as:

$$L = \ln (p / 1-p) = B_0 + B_1x \quad \text{eq. 1-1}$$

where p is the proportion or probability of a given outcome in a population, x is an explanatory variable, and L is the natural logarithm of the odds of that outcome in the population.

The model can be extended to include “ n ” number of explanatory variables, such that:

$$L = \ln (p / 1-p) = B_0 + B_1x + \dots + B_nx \quad \text{eq. 1-2}$$

In **Eq.1-1**, we can define the left side “ $\ln (p / 1-p)$ ” logit or log-odds function, and “ $p / 1-p$ ” is the odds.

Here, odds signify the ratio of probability of success to probability of failure. Therefore, in Logistic Regression, linear combination of inputs are mapped to the log(odds), as expressed in eq-1-1 and eq-1-2.

Let us better understand this model by using this model in in an **example**:

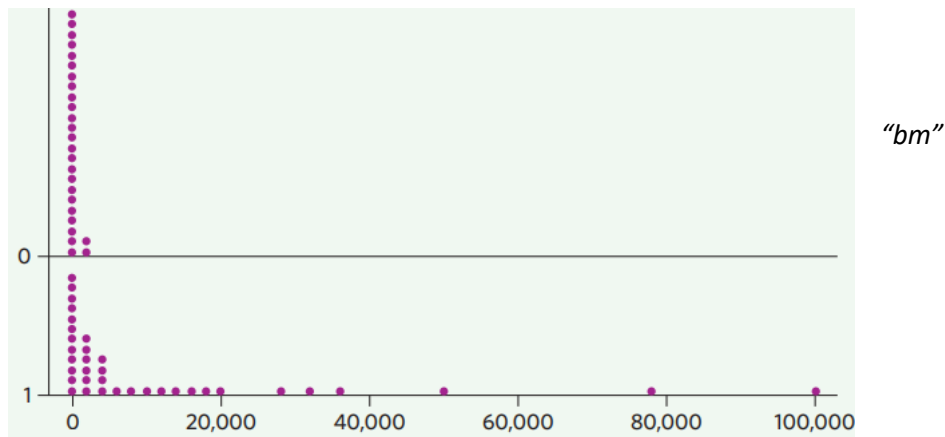
When considering the etiology of meningitis, many variables are examined. Let us consider for this example that meningitis – an inflammation of the outer membrane protecting the brain – can be caused by either viral or bacterial infection. We understand that meningitis is potentially deadly and must be treated promptly.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

For this examine we will evaluate a research study, where test results from 352 patients with acute meningitis were evaluate. These patients were later unambiguously diagnosed with either viral or bacterial meningitis. Using this set of data, we can define the variable Bacterial Meningitis (variable bm) as having a binary response (variable y). $Y=1$ when the infection is bacterial and $Y=0$ when the infection is viral.

Furthermore, immune response to the infection was assessed using white cell count per mm^3 of CSF (variable $wcsf$). Looking at the data set, we can evaluate if there is a model that would help predict whether a case of acute meningitis is viral or bacterial.

First step is to visualize the distribution of $wcsf$ in viral and bacterial cases using a dot plot. The two plots are stacked to share a common axis, making the comparison easier. All individuals who have a high white cell count had bacterial meningitis. It is clear that there is some kind of relationship between $wcsf$ and the etiology of meningitis.



White cell count in the CSF (per mm^3) " $wcsf$ "

Using the data set and from the dot plot, we must now estimate regression coefficients. There can be infinite sets of regression coefficients. The maximum likelihood estimate is that set of regression coefficients for which the probability of getting the data we have observed is maximum. If we have binary data, the probability of each outcome is simply π if it was a success, and $1-\pi$ otherwise. Therefore, we have the *likelihood* function:

$$\mathcal{L}(\beta; \mathbf{y}) = \prod_{i=1}^N \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i)$$

eq. 1-3

To determine the value of all coefficients, log of *likelihood* function is taken (doing so, does not change the properties of the function). Next, this $\log(\text{likelihood})$ is differentiated and using iterative techniques like Newton-Raphson method or Gradient descent method (this can be very math heavy and this step is typically completed on a computer software designed for that purpose), values of parameters that maximize the $\log(\text{likelihood})$ are determined.

For our example, we find that $B_0 = -1.193$ (constant) and $B_1 = 0.0018$ (wcsf). We can now construct the logistic regression model with coefficients as:

$$L = \ln(p / (1-p)) = -1.193 + 0.0018x \quad \text{eq. 1-4}$$

This regression has basically converted a sigmoidal “S-Shaped” curve to a linear relationship.

Note also that the logistic model does not represent values of p that are either 0 or 1. Instead, it provides models in which p can come arbitrarily close to 0 or 1. And we can use a little math to derive the equation in terms of the probability “ P ”:

$$p = e^{B_0+B_1x} / (1 + e^{B_0+B_1x}) \quad \text{eq. 1-5}$$

In our example, we can use eq 1-5 and apply the variables B_0 and B_1 to create a probability of table:

x (wcsf)	0	100	500	1000	10000
$p(\text{bacterial})$	0.2321	0.2657	0.4265	0.6465	0.9999

and establish that there is indeed an association between the white cell count in the CSF and the source of acute meningitis. We can now use this to help guide future cases where the CSF white blood count is known and help predict the probability of bacterial vs viral meningitis.

Given that logistic regression uses past experience of a group of patients to estimate the odds of an outcome by modeling or simulating that experience, let us now consider a hypothetical patient in whom meningitis is suspected, and you want to determine the probability that antibiotics are needed.

We can see that around a CSF white count of 663 the probability of bacterial meningitis is 50% and at a white count of 0, the probability is 23%. This is likely because *probability* is constrained between 0 and 1 and *odds* are constrained between 0 and infinity. The importance of this is that a large odds ratio (OR) can represent a small probability and vice-versa. Therefore, to translate our model into clinical practice we can have to identify a **reference point**. As an example, a reference point of 25% would warrant antibiotic use if the CSF white blood count is over 52. In most instances, clinicians can improve validation – or predictive accuracy – by using a large data set to create the model.

Moreover, an increase of 1 white cell is not medically relevant. However, when the explanatory variable in a logistic regression model is categorical, the odds ratio is commonly used to compare the odds between two conditions.

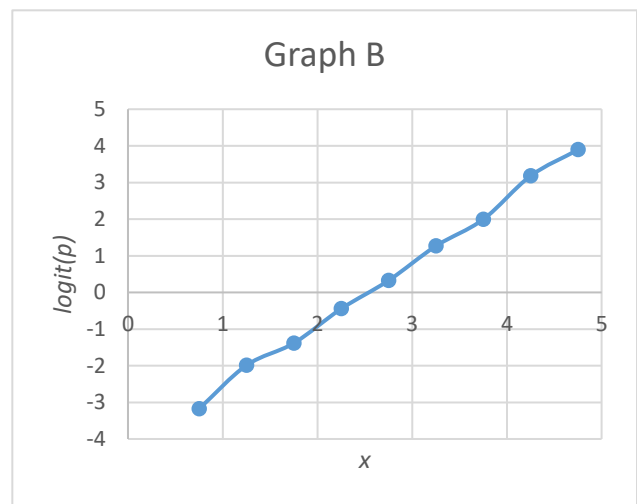
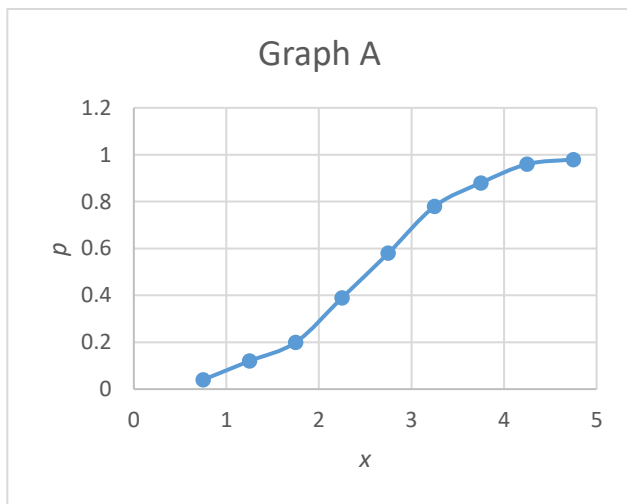
Let us look at another example:

Consider a sample of 2000 patients whose levels of a metabolic marker have been measured. Here we will evaluate how death (1) vs survival (0) can be predicted by the level of hypothetical metabolic marker. To make it easier to visualize this data set, we will create metabolic marker “groups”.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Metabolic marker level (x)	# of Patients (pts)	# of Deaths (dth)	Proportion of Death ($p = dth/pts$)
0.5 to <1.0	182	7	0.04
1.0 to <1.5	233	27	0.12
1.5 to <2.0	224	44	0.20
2.0 to <2.5	236	91	0.39
2.5 to <3.0	225	130	0.58
3.0 to <3.5	215	168	0.78
3.5 to <4.0	221	194	0.88
4.0 to <4.5	200	191	0.96
>= 4.5	264	260	0.98
Total	2000	1112	

Graph A suggests that the probability of death increases with the metabolic marker level. The relationship is nonlinear and that the probability of death changes very little at the high or low extremes of marker level. This pattern is typical because proportions cannot lie outside the range from 0 to 1. The relationship can be described as following a sigmoid-shaped curve. **Graph B** shows the logit-transformed proportions and is fairly linear. The relationship between probability of death and marker level x could



therefore be modelled using the $\text{logit}(p) = B_0 + B_1x$ equation.

Once again using “Maximum Likelihood Estimation”, we can derive the coefficient $B_1 = 1.690$ and $B_0 = -4.229$. Moreover, we can also derive that the Odds Ratio for each 1 unit increase in the value of X is $OR = e^{B_1}$ (or $e^{1.690} = 5.4$ in our example). Lastly, as derived in eq-1.5, predicted probability of death for any given value of metabolic marker in this example will be:

$$p = \frac{e^{(-4.229+1.69x)}}{1 + e^{(-4.229+1.69x)}}$$

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

For its clinical utility, we apply these equations to arrive at the following conclusions:

- Predicted probability of death at metabolic marker level of 2.0 = 0.3.
- Predicted probability of death at metabolic marker level of 3.0 = 0.7
- Predicted value of marker when predicted probability equals 0.5 – that is, at which the two possible outcomes are equally likely. $X = 2.5$. This can be considered our goal
- The odds of death for a patient with a marker level of (“x+1”) 3.0 is 5.4 times that of a patient with marker level (“x”) 2.0.

This model can be used to set goals of therapy to reduce metabolic marker below certain points and help in patient education and decision making.

References:

1. Spanos, F. E. Harrell, and D. T. Durack. Differential diagnosis of acute meningitis: an analysis of the predictive value of initial observations. *Journal of the American Medical Association*, 262 (1989), pp. 2700–2707.
2. Hosmer DW, Lemeshow S: *Applied Logistic Regression*. 2nd Ed. John Wiley and Sons; 2000.
3. Bewick V, Cheek L, Ball J. *Statistics review 7: Correlation and regression*. Crit Care. 2003
4. Meurer WJ, Tolles J. Logistic Regression Diagnostics: Understanding How Well a Model Predicts Outcomes. *JAMA*. 2017;317(10):1068–1069. doi:https://doi.org/10.1001/jama.2016.20441

2018 and Jan 2020

IV.9 Introduction to Propensity Score Matching (Jiazuo Henry Feng)

Introduction

In the field of research, one of the most reliable forms of clinical study is the double-blinded randomized clinical trial. However, these studies require significant resources, time, and patient volume for proper selection and consent in order for the study to be completed. In addition, ethical barriers with randomized human studies restrict certain hypotheses to be tested. Thus, some observational data cannot be subject to random hypothesis testing and are bound by the controls set by the study parameters.

The concept of propensity matching is better shown through asking a simple question:

Do hospital-acquired clostridium difficile infections add burden to patients and hospital cost?

Consider a hospital system interested in the question above. A simple formulation of the hypothesis would be, for example- does hospital acquired clostridium difficile (c-diff) infection increase mortality among inpatients and increase hospital costs? However, testing this hypothesis runs into a major ethical barrier - choosing which patients to have c-diff infections. Such a randomized clinical trial would never pass the IRB!

Therefore, we are limited to observational, retrospective data. It is also safe to assume that those infected received standard of care therapy. How then do we compare a group of patients that were c-diff infected to one which was not c-diff infected? In addition, without prospective controls (as there would be in an RCT), how are we to compare the myriad of covariates in our patient samples?

Observational, retrospective data thus inherently cannot reliably answer the hypothesis question but propensity score matching attempts to get close to an answer.

Note that this guide is not to teach the technical aspects of how to run a propensity score algorithm on a computer, but rather the broad concepts that feed into the algorithms.

Step 1: Data gathering

When thinking about PSM, data gathered from the electronic records should be inclusive of the covariates you wish to analyze. Keep in mind that your set must be such that your covariates reduce your analyses' selection bias (a common pitfall of most propensity matching analyses). However, the more covariates you try and include in an analysis, the longer your analysis will run and the more likely you will have incomplete matching in the end. There is a delicate balance!

Step 2: Data analysis

PSM requires a good understanding of logistic regression. Please refer to the logistic regression section of this guide. Logistic regression is used to general the propensity score, or

$$P(\text{outcome} | x_i)$$

This logistic model is generic, which is the representation of the probability of an outcome (P) based on certain variables, x_i .

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

In your dataset, you would have split individuals into 2 sets - those who have had a specific “outcome” (also can be termed the “treatment” arm) and those who did not (this would be your control group). When you analyze each specific individual in the set, you will perform a logistic regression analysis on the probability or *propensity* that the outcome could happen with the selected covariates.

Let’s go back to the clinical study example - we have 2 outcomes groups: individuals that had cdiff and those who did not. Hence, we can label those with c-diff as $T=1$ and those who were not infected as $T=0$. Next, we would go through every single individual and calculate a specific propensity score based on the other covariates that we selected to analyze, x_i , such as age, gender, antibiotic administration, etc. Thus, for each individual, we can generate a regression coefficient - a number representing the “best-fit” model, which represents our propensity score for each individual.

Step 3 - Matching

Now that you have estimated propensity scores for all of your individuals in your dataset, you can now proceed with matching your positive outcomes group (treatment group) and negative outcomes group (control group).

There are different matching algorithms in score matching. The most popular is bipartite matching, which involves a matching ratio and a matching algorithm (in a table below).

Ratio	Definition
One-to-one	One treatment individual to one control individual
Variable	Allows the algorithm to decide the matching ratio and thus can generate 1:1, 1:2, 1:n matching pairs for optimal matching
Fixed	Each control is matched to a specified number of treatment individuals

Algorithm	Definition
Greedy	Sets allowable “distance” or absolute value between propensity scores
Nearest neighbor	Matches each treatment group individual with closest possible control individual

There are specific software available for matching, including Excel, R, SAS, and Strata.

Your matched set will be the dataset that you analyze. Back now to our example - now that you have c-diff patients matched with non-c-diff patients matched based on similar characteristics of covariates, you may begin to analyze this new dataset for correlations with mortality and hospital cost. In essence, what you have done is gone into your raw dataset pool, fished out patients in that pool that have similar characteristics as the treatment option, and to the best of your ability eliminated possible selection biases. In other words, without doing a randomized trial, you've generated a very similar control group to your treatment group, and now can compare this new control group to the treatment (or outcomes) group!

Step 4: Evaluate for validity

Now that you have your matched set, make sure that the covariates in your groups are appropriately distributed between the treated and control groups, i.e., there are statistically insignificant differences in your matched groups (similar to how researchers present baseline statistics of their study groups and show a non-significant p-value to prove that the experimental and control groups are similar).

CONCLUSION

The propensity score matching method is an effective tool in medicine. It allows the user to choose patients in a dataset that are equal in quality, as a function of covariates. This is almost as good as a randomized trial, as it almost completely eliminates selection biases. However, nothing is as good as real life - selection biases can still occur given inappropriate or insufficient covariates analyzed. In other words, one is still limited in how many covariates one can measure *in silico*. If too few covariates are analyzed, one runs the risk of having many biases in the final matching and analysis steps. If too many covariates are analyzed, the computer would take an extremely long time to process each covariate. The lack of prospective randomizations is difficult to completely overcome.

References:

1. Infection Control & Hospital Epidemiology. Volume 34, Issue 6 June 2013 , pp. 588-596
2. Rosenbaum, Paul R.; Rubin, Donald B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika*. 70 (1): 41–55
3. https://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departamental_units/mchp/protocol/media/propensity_score_matching.pdf

Helpful websites:

Youtube search - propensity score matching

August 2018

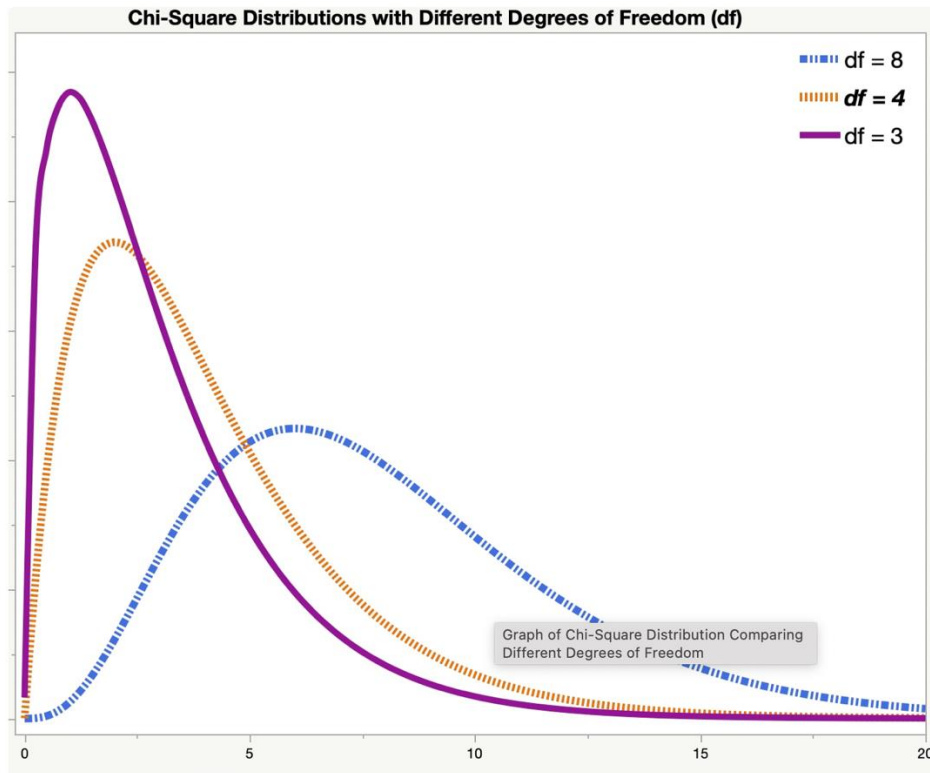
IV.10 What is Chi-square (X^2) testing? (Lauren Bernal, GSM4)

What is a Chi-square test?

A Chi-square (X^2) test is a hypothesis testing model. Two common Chi-square tests that are used are the ‘goodness of fit’ and the ‘test of independence’. The Chi-square test of independence measures the association between two categorical variables. The Chi-square goodness of fit test is used when you are trying to determine whether your data is as expected; it is often used to evaluate whether the sample data is representative of the population. The concept is that you compare the observed data from your study/experiment against the expected values if the null hypothesis is true. ¹

	Chi-Square Goodness of Fit	Chi-Square Test of Independence
Number of variables	One	Two
Purpose	Determines if sample data matches a population; fits one categorical variable to a distribution	Compares two variables in a contingency table and determines if there is a relationship
Degrees of freedom	Number of categories minus 1	(# of categories for first variable minus 1)x(# of categories for second variable minus 1)

An important component of the Chi-square test are the degrees of freedom. The mathematical definition of degrees of freedom is the rank of a quadratic form, which when translated from mathematical speak is that each item being estimated requires the consumption of ONE degree of freedom, and the remaining degrees are used to estimate variability. Increasing the number of degrees of freedom leads to an increase in the mean of the distribution as well as the probability density of larger values (see below graph of three density functions, referenced from JMP). Therefore, the higher the degrees of freedom, the more closely that the Chi-square distribution looks like a normal distribution. ^{2,7}



(From Ref. 1)

What is the Chi-square NOT used for?

A Chi-square test is meant to test if data is as expected and the probability of independence of a distribution of that data. A Chi square test, however, will NOT give any details about the relationship between the data. Once you have determined the probability that the two variables are related using the Chi-squared test, you can use other statistical models to explore the relationship between the two variables.

How to perform a Chi-square test

- a. Define null and alternative hypotheses before collecting data
- b. Decide on the alpha value.
 - i. I.e., setting $\alpha=0.05$ when testing for independence means that you have a 5% risk of concluding two variables are independent when they are not.
- c. Have data values that are a simple random sample from the population (data set that is large enough so that at least five values are expected in each of the observed data categories)
- d. Calculate test statistic using the formula:

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared

O_i = observed value

E_i = expected value

- e. Identify the degrees of freedom
- f. In a Chi-square table identify the cell corresponding to your pre-assigned alpha level and degrees of freedom; the value in the corresponding cell gives you the Chi-square distribution value.^{3,4}

Use the blank space below for your own questions or calculations:

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

(From Ref. 5)

- g. Interpret your test statistic to the distribution value from the table
 - i. Calculated test statistic < Chi-square value from table: Fail to reject the null hypothesis.
 - ii. $X^2 = 0$ means that the observed and expected values were equal, so there is no difference
 - iii. Calculated test statistic > Chi-square value from table: Reject the null hypothesis

Let's run through this process with an imagined and then a more realistic example: Is endometrial cancer more likely to be diagnosed in patients who are in a low-income bracket/socioeconomic status? (NOTE: the following numbers are made up and only used to illustrate an example; this is NOT information from a research study. Socioeconomic status in a true study would identify the cutoff points for each category but is not defined in this hypothetical example).

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

H_0 : The frequency of endometrial cancer diagnosis is not expected to be different based on a patient's socioeconomic status.

- Expected frequency: High = Moderate = Low

H_a : The frequency of endometrial cancer diagnosis is different based on a patient's socioeconomic status.

Hypothetically data on patients diagnosed with endometrial cancer was collected from multiple sites in the U.S. Number of patients (N) = 1500. Degrees of freedom = 2, $\alpha = 0.05$.

Goodness of Fit Example:

	Observed Frequency	Expected Frequency	(Obs-Exp) ² /Exp
High Socioeconomic status	50	500	405
Moderate Socioeconomic status	200	500	180
Low Socioeconomic status	1250	500	1125

$$X^2 = 405 + 180 + 1125$$

$$X^2 = 1710$$

The X^2 critical value = 5.991 (per above z chart) when degrees of freedom = 2 and the $\alpha = 0.05$.

Therefore, our X^2 value > the critical value assigned by the table, meaning that we reject our null hypothesis. In this example, that would indicate that the distribution of endometrial cancer patients is not equally distributed across socioeconomic classes. Therefore, further studies could then examine the relationship/association between socioeconomic status groups and endometrial cancer; for example, by looking at patients across different socioeconomic statuses and seeing if they have been diagnosed with endometrial cancer.

(NOTE AGAIN: the above example and numbers are not real.)

Now let's look at a recently published trial and use this as an example from a RCT: In patients with a history of a failed vaginal cerclage does the placement of an abdominal cerclage or a high vaginal cerclage offer greater benefit than traditional low vaginal cerclage?

H_0 : Abdominal cerclage is not associated with improved outcomes in comparison to traditional low vaginal cerclage for patients with previous failed vaginal cerclage.

H_a : Abdominal cerclage is associated with improved outcomes in comparison to traditional low vaginal cerclage for patients with previous failed vaginal cerclage.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

	Delivery at <32 weeks completed gestation	Delivery after 32 weeks	Total
Abdominal cerclage	3	36	39
Traditional low vaginal cerclage	11	22	33
Total	14	58	72

Expected Frequency = (row sum x column sum)/table sum

Expected frequency of delivery <32 weeks with abdominal cerclage = $(39 \times 14) / 72 = 7.58$

Expected frequency of delivery <32 weeks with low vaginal cerclage = $(33 \times 14) / 72 = 6.42$

$$\chi^2 = [(3-7.58)^2/7.58] + [(11-6.42)^2/6.42]$$

$$\chi^2 = 2.77 + 3.27$$

$$\chi^2 = 6.04$$

In this scenario with one degree of freedom, the calculated χ^2 of 6.04 is > than the critical value of 3.841 at an alpha of 0.05. Using the EBM calculator (<https://ebm-tools.knowledgetranslation.net/calculator>) when analyzing this study, the calculated Chi-square value is 5.955 with a p-value of 0.015. Variations in calculations could possibly be explained by rounding of significant figures. In a χ^2 analysis, the p-value is the probability of obtaining a χ^2 >= to the current experiment. In other words, it is the probability of deviations from what is expected due to mere chance. So in this example, our χ^2 value is 6.04, which falls between an alpha of 0.025 and 0.01. Thus, our χ^2 could be said to be due to chance between 1% to 2.5% of the time. The significance level was set at 0.05 for this study, or saying the level at which a 5% chance of our χ^2 being due to chance is acceptable and statistically significant. Our χ^2 value is statistically significant with a small % likely due to chance. Therefore, we reject the null hypothesis that abdominal cerclage is not associated with improved outcomes in comparison to traditional low vaginal cerclage.

Further, statistical analysis such as calculating the Relative Risk Reduction (RRR) and Number Needed to Treat (NNT), gives clinical meaning to the observed statistically significant difference that we see with the above χ^2 analysis.

For this example:

$$\text{RRR} = 77\% \text{ with } 95\% \text{ CI } (24.2 \text{ to } 93)$$

$$\text{NNT} = 4 \text{ with } 95\% \text{ CI } (14 \text{ to } 2)$$

These demonstrate a strong risk reduction with a low NNT of preterm birth at <32 weeks in women with a history of failed vaginal cerclage with the use of an abdominal cerclage. (Article reference: Am J Obstet Gynecol. 2020 Mar;222(3):261.e1-261.e9. PMID:31585096)

References

1. *The chi-square test*. JMP. (n.d.). Retrieved November 12, 2021, from https://www.jmp.com/en_us/statistics-knowledge-portal/chi-square-test.html.

2. Dallal, G. E. (2020, December 20). *Degrees of Freedom*. The Little Handbook of Statistical Practice. Retrieved November 5, 2021, from <http://www.jerrydallal.com/LHSP/dof.htm>.
3. Stephanie Glen. "Chi-Square Statistic: How to Calculate It / Distribution" From StatisticsHowTo.com: Elementary Statistics for the rest of us!
<https://www.statisticshowto.com/probability-and-statistics/chi-square/>
4. *SPSS tutorials: Chi-Square test of Independence*. LibGuides. (n.d.). Retrieved November 5, 2021, from <https://libguides.library.kent.edu/spss/chisquare>.
5. *Chi-square table*. Z Score Table. (n.d.). Retrieved November 5, 2021, from <http://www.z-table.com/chi-square-table.html>.
6. Frieden B.R. (2001) The Chi-Square Test of Significance. In: Probability, Statistical Optics, and Data Testing. Springer Series in Information Sciences, vol 10. Springer, Berlin, Heidelberg. https://doi-org.dartmouth.idm.oclc.org/10.1007/978-3-642-56699-8_11
7. Taboga, Marco (2017). "Chi-square distribution", Lectures on probability theory and mathematical statistics, Third edition. Kindle Direct Publishing. Online appendix.
<https://www.statlect.com/probability-distributions/chi-square-distribution>.

Submitted December 2021

IV.11 One-Way Analysis of Variance (ANOVA)—what is it, when is it used, and sample calculation (Marie Syku, GSM4)

What is ANOVA?

Analysis of variance (ANOVA) is a tool used in statistics to determine differences between research results from three or more unrelated groups/samples. This is done by the examination of variance within each group and the variance between groups to determine if observed differences between groups are due to actual effects or random variability. As a widely used statistical method in medical research, ANOVA allows researchers to compare the effectiveness of studied treatments and interventions.

ANOVA has the flexibility to cover many experimental designs. There are different types of ANOVA, each suited for specific study designs and hypotheses, as well as the nature of the data being analyzed.

1. One-way ANOVA: Compares the mean of 3 or more independent groups to determine if there is a statistically significant difference between the groups. An example of this could be investigating the effect of three different drugs (independent variables) on glucose concentrations (dependent variable) in the blood.
2. Two-way ANOVA: Analyzes the effects of 2 independent categorical variables (factors) on a continuous dependent variable. An example of this could be evaluating if there is an interaction

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

between physical activity (low/moderate/high) and gender (male/female) on blood cholesterol levels in adolescents.

3. Repeated Measures ANOVA: analyzes the data collected from the same subjects at multiple time points or under different conditions. An example of this could be evaluating the effect of a 6-month exercise program on blood pressure in the same individual across three separate time points (pre-workout intervention, mid-way, post-intervention).

For the purposes of this chapter, we will be focusing on one-way ANOVA testing.

Minimizing Type 1 Error?

While T-tests are used to assess for significant differences between the means of *two* groups, they are not used when the number of groups exceeds two. This is because when each group is paired with another to attempt *three* paired comparisons of group means, Type I error increases. In other words, there is an increased probability of obtaining a false positive, rejecting the null hypothesis and concluding that the alternative hypothesis has significance (despite there being no real significant difference). As the number of group comparisons increases, the probability of rejecting the entire null hypothesis and obtaining a false positive also increases (Table 1).

Number of comparisons	Significance level
1	0.05
2	0.098
3	0.143
4	0.185
5	0.226
6	0.265

Table 1.

ANOVA avoids the issue of type I error inflation that commonly occurs when one attempts to compare the means of three or more groups and keeps the error rate at the alpha level that is set (typically 0.05).

The basic principle of 1-way ANOVA involves calculating F-statistics by dividing the variance *between* groups by the variance *within* groups.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SSW = \sum_{j=1}^k \sum_{i=1}^l (x - \bar{x}_j)^2$	$df_w = k - 1$	$MSW = \frac{SSW}{df_w}$	$F = \frac{MSB}{MSW}$
Between	$SSB = \sum_{j=1}^k (\bar{x}_j - \bar{x})^2$	$df_b = n - k$	$MSB = \frac{SSB}{df_b}$	
Total	$SST = \sum_{j=1}^n (\bar{x}_j - \bar{x})^2$	$df_t = n - 1$		

F = Anova Coefficient
 MSB = Mean sum of squares between the groups
 MSW = Mean sum of squares within the groups
 MSE = Mean sum of squares due to error
 SST = total Sum of squares
 p = Total number of populations
 n = The total number of samples in a population
 SSW = Sum of squares within the groups
 SSB = Sum of squares between the groups

(Table taken from reference 4)

If we assume the null hypothesis to be true (i.e. there is no significant difference in the means across groups), then the variability attributable to between-group differences should be relatively low. Most of the observed variability should be attributable to within-group differences. Consequently, the F-value will be low. If the F-value is large enough to surpass a critical threshold (which depends on the degrees of freedom and the chosen alpha level) then the associated P value will be less than alpha, the null hypothesis is rejected, suggesting a significant difference between at least two group means.

Let's do an example to illustrate this concept. Please note that the illustration below is just an example (modified from reference 6) and does not represent real-life study data:

Suppose that a pharmaceutical company wants to conduct an experiment to test out the efficacy of a brand new cholesterol lowering medication. A total of 15 random participants are selected from a larger population and assigned to one of three groups. Group 1 participants receive 0 mg/day of the medication; Group 2 receives 50 mg/day; Group 3 receives 100 mg/day. After 1 month post-treatment initiation, measurements are taken of each participant's cholesterol level. Results appear in the table below:

Group 1 (0 mg/day)	Dosage	
	Group 2 (50 mg/day)	Group 3 (100 mg/day)
210	210	180
240	240	210
270	240	210
270	270	210
300	270	240

Before we use 1-way ANOVA to investigate for any significant differences in mean cholesterol levels across the three treatment doses, we must ensure that ANOVA is the correct statistical method to use (i.e. the experimental design is compatible with 1-way ANOVA and key assumptions are met).

What assumptions does the ANOVA test make?

To conduct hypothesis testing in ANOVA, three main assumptions are made. Firstly, it is assumed that in the population, the dependent variable scores are normally distributed within each of the treatment groups. Secondly, the test assumes homogeneity of variances; namely, the variance of dependent variable scores across the different groups is equal. Finally, the dependent variable score of each group is independent of that for any other group. If these assumptions are not met, the study can lose a considerable amount of power, potentially requiring transformation of data.

In our example, the experimental design (randomization of participants) is compatible with 1-way ANOVA, and all three key assumptions have been satisfied.

Let us now return to our example:

To proceed with application of 1-way ANOVA, the following steps are taken:

1. Specify a mathematical model to describe causal factors that affect the dependent variable.
2. Specify the hypothesis to be tested.
3. Specify a significance level for the hypothesis test.
4. Calculate the total mean and the mean values for each group.
5. Calculate the sum of squares for each effect in the model.
6. Identify the degrees of freedom associated with each effect in the model.
7. Calculate the mean squares for each effect in the model based on the sums of squares and degrees of freedom.
8. Calculate the test statistics, based on the observed mean squares and their expected values.
9. Identify the P value for the test statistic.
10. Accept or reject the null hypothesis, based on the P value and significance level.

STEP 1)

For our mathematical model, we can use the following:

$$X_{ij} = \mu + \beta_j + \varepsilon_{i(j)}$$

where X_{ij} is the cholesterol level for subject i in treatment group j , μ is the population mean, β_j is the effect of the dosage level administered to subjects in group j ; and $\varepsilon_{i(j)}$ is the effect of all other extraneous variables on subject i in treatment j .

STEP 2)

For our null hypothesis we can state the following: the dosage level of medication has no effect on the cholesterol level in any of the treatment groups. If the null hypothesis is true, the mean cholesterol level across treatment groups should be equal.

For our alternative hypothesis we can state the following: the dosage level does have an effect on the cholesterol level *in at least one* treatment group. If the alternative hypothesis is true, at least one pair of mean cholesterol scores across treatment groups should be unequal.

STEP 3)

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

The significance level (alpha or α) is typically specified by the study investigators. For our purposes, let us choose a significance level of 0.05.

STEP 4)

Calculating grand mean and group means-

We can compute the grand mean (\bar{X}) and group means as follows:

$$\bar{X} = (1 / 15) * (210 + 210 + \dots + 270 + 240)$$

$$\bar{X} = 238 = \text{grand mean}$$

$$\bar{X}_j = (1 / n_j) \sum_{i=1}^{n_j} (X_{ij})$$

where n_j is the sample size in Group j

$$\bar{X}_1 = 258$$

$$\bar{X}_2 = 246$$

$$\bar{X}_3 = 210$$

STEP 5)

Calculating sum of squares. A sum of squares is the sum of the squared deviations from a mean score.

1-way ANOVA uses three sums of squares:

1. Between-group sum of squares (SSB)- measures the variation of group means around the grand mean. Can be calculated through the following formula:

$$SSB = \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{X}_j - \bar{X})^2 = \sum_{j=1}^k n_j (\bar{X}_j - \bar{X})^2$$

$$SSB = 5 * [(238-258)^2 + (238-246)^2 + (238-210)^2]$$

$$SSB = 6240$$

2. Within-group sum of squares (SSW)- measures variation of all scores around their respective group means. Can be calculated through the following formula:

$$SSW = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

$$SSW = 2304 + \dots + 900 = 9000$$

3. Total sum of squares (SST)- measure variation of all scores around the grand mean. Can be calculated from the following formula:

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

$$SST = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

$$SST = 784 + 4 + 1084 + \dots + 784 + 784 + 4$$

$$SST = 15,240$$

From our calculations you can see that the total sum of squares is equal to the between-groups sum of squares plus the within-groups sum of squares ($SST = SSB + SSW$; $15,240 = 6240 + 9000$)

STEP 6)

Degrees of freedom (df) equals the number of *independent sample points* used to calculate a statistic minus the number of *parameters estimated* from the sample points.

As examples, let's identify the degrees of freedom associated with the various sum of squares calculations we did above.

1. Between-group degrees of freedom

$$SSB = \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{X}_j - \bar{X})^2 = \sum_{j=1}^k n_j (\bar{X}_j - \bar{X})^2$$

Here, the formula uses k independent sample points (sample means, \bar{X}_j) and 1 parameter estimate (the grand mean \bar{X} which was estimated more the sample points). So the between-groups sum of squares has $k-1$ degrees of freedom = $df_{BG} = k - 1 = 5 - 1 = 4$.

2. Within-groups degrees of freedom

$$SSW = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

Here, the formula uses n independent sample points (the individual subject scores, X_{ij}) and k parameter estimates (the group means \bar{X}_j which were estimated from the sample points). So the within-groups sum of squares has $n-k$ degrees of freedom = $df_{WG} = 15 - 3 = 12$.

3. Total degrees of freedom:

$$SST = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

$$\begin{aligned}SST &= 784 + 4 + 1084 + \dots + 784 + 784 + 4 \\SST &= 15,240\end{aligned}$$

Here, the formula uses n independent sample points (the individual subject scores, X_{ij}) and 1 parameter estimate (the grand mean \bar{X} , which was estimated from the sample points). So the total sum of squares has $n-1$ degrees of freedom = $df_{TOT} = 15 - 1 = 14$.

STEP 7)

To calculate the mean square, an estimate of the population variance, you divide the sum of squares (SS) by its corresponding degrees of freedom (df).

$$MS = SS / df$$

To conduct 1-way ANOVA, we are interested in two mean squares: within-groups mean square and between groups mean square.

1. Within-group mean square (MS_{WG})- refers to the variation due to differences among experimental units within the same group. Can be calculated as follows:

$$MS_{WG} = SSW / df_{WG} = 9000 / 12 = 750$$

2. Between-group mean square (MS_{BG})- refers to variation due to differences among experimental units within the same group plus variation due to treatment effects. Can be calculated as follows:

- $MS_{BG} = SSB / df_{BG} = 6240 / 2 = 3120$

STEP 8)

The test statistic, F ratio, is a convenient measure that can used to test the null hypothesis. It measures the ratio between MS_{BG} and $MS_{WG} = MS_{BG} / MS_{WG} = 3120 / 750 = 4.16$.

STEP 9)

The P-value denotes the probability of obtaining a result more extreme than the observed experimental outcome, assuming the null hypothesis is true. In 1-way ANOVA, the P-value represents the probability that an F statistic would be bigger than the actual F ratio calculated from experimental data. To get the associated P-value we can use free online calculators, including Stat Trek's F Distribution Calculator (<https://stattrek.com/online-calculator/f-distribution>). Entering the between-groups and within-groups degrees of freedom (2 and 12, respectively) and the F ratio (4.16) into the calculator, we obtain $P(F \geq 4.16) = 0.04242$, yielding a P value of 0.04.

Degrees of freedom (v_1)	2
Degrees of freedom (v_2)	12
f Statistic (f)	4.16
Probability: $P(F \leq 4.16)$	0.95758
Probability: $P(F \geq 4.16)$	0.04242

Calculate

STEP 10)

Based on the calculated P value of 0.04, which is less than the 0.05 significance level set earlier in the experiment, we can reject the null hypothesis and conclude that the mean cholesterol level in at least

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

one of the treatment groups was significantly different from the mean cholesterol level in another group.

Limitations of ANOVA and post-hoc test?

Conclusions that can be derived from ANOVA testing do not come without limitations. When the null hypothesis gets *rejected*, it suggests that the means of the three groups may differ and at least one group may show a difference. However, one-way ANOVA cannot tell you which specific groups were significantly different from each other—only that at least two groups were. This necessitates an additional process of verifying through post-hoc analysis. One of the most well-known methods is the Bonferroni's correction. Briefly, the significance level is divided by the number of comparisons and applied to the comparisons of each group. As an example, when comparing population means of three independent groups A, B and C at a significance level of 0.05, the significance level for comparisons of one group to another (A+B, A+C, B+C) would be $0.05/3 = 0.017$. Additional post-hoc tests that can be used include Turkey, Scheffe, and Holm methods. More in-depth information on the Bonferroni's correction and post-hoc analysis can be found in other sections of the EBM guide.

References:

1. St, Lars, and Svante Wold. "Analysis of variance (ANOVA)." *Chemometrics and intelligent laboratory systems* 6.4 (1989): 259-272.
2. Foster, G., Lane, D., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H. (2022, May 13). *11.4: Anova and type I error*. Statistics LibreTexts. [https://stats.libretexts.org/Bookshelves/Applied_Statistics/An_Introduction_to_Psychological_Statistics_\(Foster_et_al.\)/11%3A_Analysis_of_Variance/11.04%3A_ANOVA_and_Type_I_Error](https://stats.libretexts.org/Bookshelves/Applied_Statistics/An_Introduction_to_Psychological_Statistics_(Foster_et_al.)/11%3A_Analysis_of_Variance/11.04%3A_ANOVA_and_Type_I_Error)
3. Kim TK. Understanding one-way ANOVA using conceptual figures. *Korean journal of anesthesiology*. 2017 Feb 1;70(1):22-6.
4. Admin. (2022, January 3). *ANOVA formula in statistics with solved example*. BYJUS. <https://byjus.com/anova-formula/>
5. Schober, Patrick, and Thomas R. Vetter. "Analysis of variance in medical research." *Anesthesia & Analgesia* 131.2 (2020): 508-509.
6. Berman H.B., "One-Way Analysis of Variance: Example", [online] Available at: <https://stattrek.com/anova/completely-randomized/one-way-example>

Submitted 12-4-2023

IV.12 Post Hoc Analysis - What, Why, How, and What to Worry About? (Linda Morris, GSM4)

What is it?

The Latin phrase “post hoc” means “after this” and “pre hoc,” therefore, is “before this.” In terms of research studies, the pre hoc analysis is the one incorporated into the experimental design. This is the basis of the scientific method, when a specific hypothesis is proposed, tested by an experiment, analysis is determined *before* the experiment is run, and the results either support or reject the null hypothesis. In clinical trials, the pre hoc analyses are the primary and secondary outcomes named in the study design. On the contrary, the post hoc analysis is any test subsequently run on the observed data after the experiment is complete. The post hoc analysis allows you to fit a hypothesis to an observed result, rather than test a specific hypothesis.

Why do we use it?

The main benefit of post-hoc analyses is that they have the ability to reveal patterns in the data that were not the primary objective of the study. This can be advantageous, particularly in very exploratory experimental fields. For instance, when investigating the effects of a newly developed drug, the post hoc analysis can point to previously unknown uses and particularly impacted groups that were not predicted by the initial study question. These results allow us to observe possible relationships and craft new hypotheses based on them. The newly discovered statistical relationships found in a post hoc analysis can suggest cause and effect relationships and indicate distinct clinical phenotypes in complex diseases. These new hypotheses can then be tested in additional clinical trials via pre-hoc methods to assess their reproducibility and validity.

Post hoc analyses have also become more common with the growth of large data registries. If the wealth of information from large multicenter clinical trials that is stored in clinical registries was only examined via the planned pre-hoc analysis, then we would waste the opportunity to further explore a vast amount of information, as well as the time and resources spent collecting it. When used correctly, post hoc analyses can help us better understand trial results, the population studied, and the direction in which to steer new research.

How do we use it?

The most common post hoc test is probably the Bonferroni Procedure. This test is a post hoc multiple-comparison correction. This method allows several variables to be analyzed, while limiting data falsely appearing statistically significant.

Some of the other common post hoc tests which also attempt to limit falsely significant results are:

- Duncan’s new multiple range test (MRT)
- Dunn’s Multiple Comparison Test

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

- Fisher's Least Significant Difference (LSD)
- Holm-Bonferroni Procedure
- Newman-Keuls
- Rodger's Method
- Scheffé's Method
- Tukey's Test
- Dunnett's correction
- Benjamini-Hochberg (BH) procedure

What to worry about?

So why do all those tests exist to limit falsely significant results in post hoc analyses? One of the main issues with post hoc analysis is that significant results will occur by chance if you perform numerous tests. The family-wise error rate (FWER) is a way to describe this phenomenon, and means "the probability of making one or more false discoveries, or type 1 errors, when performing multiple hypotheses tests."

How often do we see studies when the overall outcome was non-significant, but one or two particular subgroups are highlighted as having a significant effect? Well, it turns out we need to be careful when interpreting these seemingly significant data into account.

The reality is, if investigators do enough post-hoc analyses using different subgroups, it is nearly certain they will find something statistically significant. This problem is called "multiplicity" in statistics, and results in inflated false positives. For example, even if a clinical trial showed no true treatment effect as we talked about above, if you split the study population into 20 mutually exclusive subgroups, the probability of at least one significant but false positive result at a p-value of 0.05 is 64%. If you increase the subgroups to 60, then you can expect to find up to three statistically significant interaction tests ($p < 0.05$) on the basis of chance alone. In other words, in these situations, the *majority of the time*, some statistically significant results will occur as a result of chance.

This problem with post hoc analyses can be compounded when studies don't clarify their methods for these analyses and don't necessarily elucidate how many subgroups were examined. This allows the possibility for researchers to perform an unlimited amount of separate analyses in the hopes of finding something with a P value lower than 0.05. In this case, the investigators may then only present the few statistically significant relationships they found, and therefore, cherry pick the seemingly relevant data, which can be very misleading. This approach to analysis has been described as similar to an archer who targets at a barn and then paints a target around where the arrow hits. In the same way, post hoc analyses can end up giving the false impression of a "bull's eye." In reality, we know a target shows how accurate the shot was only if it was in place before the arrow flew. When used appropriately, the subgroup data can be useful in steering the direction of future clinical trials in order to confirm the

statistical relationship as mentioned above, but cannot be relied upon on their own, in the post-hoc form.

Another major problem that can arise from post hoc analysis is the post hoc power adjustment. A power threshold of 80% is commonly used, but a sample size large enough to achieve this can be difficult in some fields, where research budgets or rare conditions can make large study populations infeasible. A 2017 article published in the Journal of Surgical Research discusses a phenomenon where, in order to combat the sample size issue, some researchers will use post hoc power calculations with observed effect sizes to demonstrate that studies are underpowered and use this as evidence to advocate for lower power thresholds in their research, in this case, in the surgical field. However, the authors criticize this strategy and argue that lower power thresholds based on observed effect size (post hoc power adjustments) end up risking higher false positive rates and end up making it more difficult to differentiate between statistical noise and clinically meaningful effects.

An article in the Annals of Translational Medicine (ATM) provides an example of why relying too heavily on post hoc analyses has pitfalls. The article describes investigations of the drug pridopidine in Huntington's Disease (HD). The drug is a dopaminergic system modulator and has been tested in four large scale randomized controlled trials with over 1,000 participants, however, the primary trial outcome was not achieved in any of these studies. When the primary outcome of improvement in a composite cognitive score was not met in the first study, investigators relied on post hoc analysis results that pointed to a possible effect in total motor score. Three subsequent large scale RCTs were conducted with the primary outcome of investigating motor score and again, the primary and secondary experimental outcome results were not found to be significant. Subgroup analysis of the fourth trial showed that a significant effect may be present in earlier disease stages, and this was used to draw the conclusion that pridopidine may be an effective agent for HD. The ATM critics argue that investigators should consider the possibility that positive post hoc analyses may simply be wrong, and argue that we should be wary of conducting large-scale, resource-intensive investigations based on promising post hoc analyses when the overall trial and primary analysis are negative, especially when those negative results are repeated.

In Conclusion:

Overall, post hoc analyses are very important and useful when conducted correctly and put into their correct clinical and methodological context. However, care must be taken not to over interpret the results, as post hoc analyses are heavily subjected to bias, cherry-picking, and data-mining. Post hoc analyses have a key role in exploring statistical relationships in today's age of massive clinical data registries and can be vitally important to helping us discover previously non-hypothesized relationships. However, any conclusions we draw from post hoc analyses should always then be validated by high quality randomized controlled trials testing out new hypotheses using the scientific method, to find reproducible, clinically meaningful results.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

References:

1. Curran-Everett D, Milgrom H. Post-hoc data analysis: benefits and limitations. *Curr Opin Allergy Clin Immunol*. 2013 Jun;13(3):223-4. doi: 10.1097/ACI.0b013e3283609831. PMID: 23571411.
2. Fletcher J. Subgroup analyses: how to avoid being misled. *BMJ*. 2007 Jul 14;335(7610):96-7. doi: 10.1136/bmj.39265.596262.AD. PMID: 17626964; PMCID: PMC1914513.
3. Griffith KN, Feyman Y. Amplifying the Noise: The Dangers of Post Hoc Power Analyses. *J Surg Res*. 2021 Mar;259:A9-A11. doi: 10.1016/j.jss.2019.09.075. Epub 2020 Aug 22. PMID: 32843199; PMCID: PMC8211362.
4. Pamplona, Fabricio (2022-07-28). "Post Hoc Analysis: Process and types of tests". *Mind the Graph Blog*. Retrieved 2022-12-21.
5. Rodrigues FB, Ferreira JJ. The risks of converting post-hoc findings into primary outcomes in subsequent trials. *Ann Transl Med*. 2019 Dec;7(Suppl 8):S337. doi: 10.21037/atm.2019.09.105. PMID: 32016055; PMCID: PMC6976501.
6. Srinivas, Titte R; Ho, Bing; Kang, Joseph; Kaplan, Bruce. Post Hoc Analyses: After the Facts. *Transplantation* 99(1):p 17-20, January 2015. | DOI: 10.1097/TP.0000000000000581
7. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007 Nov 22;357(21):2189-94. doi: 10.1056/NEJMSr077003. PMID: 18032770.
8. Zhang Y, Hedo R, Rivera A, Rull R, Richardson S, Tu XM. Post hoc power analysis: is it an informative and meaningful analysis? *Gen Psychiatr*. 2019 Aug 8;32(4):e100069. doi: 10.1136/gpsych-2019-100069. PMID: 31552383; PMCID: PMC6738696.

Submitted 12-23-2022

IV.13 Bonferroni Correction: What is it and when to use it? (Ahmed El Hussein, GSM3)



Simply put, the Bonferroni correction is a method **used when multiple comparisons are being made in the same data set**. For example, you are testing the exam scores resulting from three different studying methods: taking notes (A) vs drawing the notes (B) vs practice questions (C). In your results, you compare method A vs B, A vs C, and B vs C for the best study method. In total, you've made 3 separate comparisons in the study (A vs B, A vs C, and B vs C). The Bonferroni correction helps you correct the P values for each comparison and avoid incorrectly rejecting the null hypothesis, causing a false positive result (type I error).

Let's dive deeper. First, let's define "**Family-Wise Error Rate**." The Family-Wise Error Rate (FWER) is the probability of making one or more false positive errors when comparing multiple groups (like in the example above) of the same data set. The equation and an example for FWER is shown below:

Family-wise error rate = $1 - (1 - \alpha)^m$, where m = the number of comparisons you're making.

Let's do an example to hit this out of the park. In the introductory example, we compared 3 different study methods: A vs B, A vs C, and B vs C. So, $n = 3$. Let's also say we set our $\alpha = 0.05$. In that case, our FWER is $1 - (1 - 0.05)^3 = 1 - 0.95^3 = 0.1426$. In other words, the probability of getting a false-positive error (type I error) on **at least one** of the hypothesis tests is over 14%. Now imagine if we did 10 comparisons. In that case, the probability of getting false positive in **at least one** of the comparisons would be 0.4013 or 40%! So, the larger the n (number of comparisons or hypothesis tests between the same sample data), the larger is your FWER.

So, now you're probably starting to see the value in the **Bonferroni correction, which is a multiple comparisons correction method**. To do the Bonferroni correction, simply follow the equation below, where m = the number of hypothesis tests being made in a set of sample data:

$$\alpha_{Bonferroni} = \frac{\alpha}{m}$$

where m is the number of tests

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

So, let's do an example again, this time with the correction applied. The Bonferroni correction yields an $\alpha_{\text{Bonferroni}} = 0.0167$. Now, if we do the equation again, we will get an $\text{FWER} = 1 - (1 - 0.0167)^3 = 1 - (1 - 0.0167)^3 = 1 - 0.9833^3 = 0.0492$ or a 4.92% false positive error rate. Using the Bonferroni method, we have reduced the FWER back down to around 5%. Now, any p-value for the comparisons above the **Bonferroni-corrected α (0.0167)** is not significant. **Only a p-value less than the Bonferroni-corrected α is significant.**

Okay, make sense? Let's conclude with a real-world research example to see it in use using the article "Daytime fluctuations of endurance performance in young soccer players: a randomized cross-over trial" by Janis Fielder, et al. This study aimed to measure differences in endurance running performance, blood lactate levels, and heart rate in young soccer players using an incremental treadmill test on two different occasions at different times **during** the day (morning versus evening).

Between groups, there was no significant difference in heart rate, lactose **concentration, running speed, or 3000-m test** after the Bonferroni correction. However, I'll specifically bring your attention to the maximum running speed in Table 1 (shown below). The original p-value for maximal lactose concentration was 0.025. This would have represented a statistically significant result in the difference in maximal running speed between morning and evening groups. However, after the Bonferroni correction was applied, the corrected p-value was 0.100. Hence, the result was no longer significant.

Let's do it together. First, we'll follow the equation for the Bonferroni correction explained above ($\alpha_{\text{Bonferroni}} = \alpha/m$). This will yield 0.05/4 (since two groups are being compared on two different occasions) and our new $\alpha_{\text{Bonferroni}}$ would be 0.0125. Any p-value obtained for comparisons between groups above that would be statistically insignificant. *It is important to note* that the Bonferroni correction can be done in another way and in this article, the authors multiplied their obtained p-value for a result by the number of comparisons being made for that result, rather than dividing the α by the number of comparisons being made and altering the significance level of the p-value. This is in effort to keep the statistically significant level of the p-value at 0.05. For example, rather than doing 0.05/4 and getting a new significance level of 0.0125 for the p-value (i.e., the results must be below 0.0125 to be significant), the authors kept the significance level of the p-value at 0.05 and multiplied whatever p-value they obtained for a comparison by the number of comparisons they made. For example, the original p-value obtained for maximum running speed was 0.025. Since the authors did 4 comparisons, they multiplied the obtained p-value of 0.025 by 4 to yield .100. The significance level in this case remained set at 0.05, yielding the result of .100 insignificant.

A similar change can be seen with the maximal lactose concentration between groups in which the result was no longer statistically significant after the Bonferroni correction was applied (i.e., after the authors multiplied the obtained p-value by 4).

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Table 1 Results for endurance running performance, blood lactate levels, and heart rate differences between morning and evening

Incremental treadmill test						
Parameter	Morning	Evening	Mean difference	corrected p-value (original)	Cohen's d (t-value)	df
Heart rate [1/min]						
Rest	86.60 (9.68)	85.73 (10.88)	0.87 (10.33)	1.00 (0.750)	− 0.09 (0.35)	14
LT	150.93 (12.13)	153.47 (10.29)	− 2.53 (9.81)	1.00 (0.334)	0.23 (− 1)	14
IAT	177.47 (7.85)	179.27 (6.31)	− 1.80 (4.87)	0.872 (0.174)	0.25 (− 1.43)	14
Max	197.13 (6.29)	198.73 (6.08)	− 1.60 (4.21)	0.814 (0.163)	0.26 (− 1.47)	14
Lactate concentration [mmol/l]						
Rest	0.84 (0.21)	0.83 (0.31)	0.01 (0.31)	1.00 (0.930)	− 0.04 (0.09)	12
LT	1.52 (0.67)	1.66 (0.61)	− 0.13 (0.40)	1.00 (0.250)	0.22 (− 1.20)	12
IAT	3.02 (0.67)	3.16 (0.61)	− 0.14 (0.40)	0.962 (0.241)	0.22 (− 2.51)	12
Max	9.15 (2.18)	10.64 (2.30)	− 1.49 (2.15)	0.110 (0.028)	0.67 (− 2.51)	12
Running speed [km/h]						
LT	8.67 (1.17)	9.00 (1.10)	− 3.30 (0.74)	0.429 (0.107)	0.29 (− 1.72)	14
IAT	11.94 (1.28)	12.12 (1.18)	− 0.19 (0.68)	1.00 (0.317)	0.15 (− 1.04)	13
Max	15.81 (1.62)	16.31 (1.59)	− 0.49 (0.76)	0.100 (0.025)	0.31 (− 2.51)	14
3,000-m test						
Time [min:sec]	12:59:00 (1:29)	13:06:00 (1:30)	− 0:07 (0:22)	0.228	0.08 (− 1.26)	15
RPE	15.31 (1.82)	15.44 (1.09)	− 0.13 (1.78)	0.783	0.09 (− 0.28)	15

Means (standard deviations) and results of the paired t-tests for daytime differences at the incremental treadmill test before the start (rest), at the onset of lactate accumulation (LT), the individual anaerobic threshold (IAT), and immediately after volitional exhaustion (max) and at the end of the 3,000-m run for time to completion (Time) and rating of perceived exertion (RPE). p-values were corrected using the **Bonferroni** method

Fiedler, Janis et al. “Daytime fluctuations of endurance performance in young soccer players: a randomized cross-over trial.” *BMC research notes* vol. 15,1 351. 24 Nov. 2022. Table 1. Results for endurance running performance, blood lactate levels, and heart rate differences between morning and evening.

Well, that’s a quick overview of the Bonferroni correction and when to use it. To summarize, the Bonferroni correction is done when you are comparing multiple groups of the same data set. In addition, the Bonferroni correction can be done by **dividing the α** by the number of comparisons being made **or by multiplying the obtained p-value** by the number of comparisons being made while keeping α untouched. Additional resources are listed below for further guidance on the use of the Bonferroni correction, as needed.

One modification of the Bonferroni correction is known as the Holm-Bonferroni Method. As shown above, the Bonferroni correction reduces the possibility of getting a statistically significant result (i.e. a Type I error) when performing multiple tests. Although the Bonferroni is simple to calculate, it suffers from a lack of **statistical power**. The Holm-Bonferroni method is also fairly simple to calculate, but it is more powerful than the single-step Bonferroni. Those interested in exploring this other way of correcting a FWER are encouraged to review a brief article by Glen.

And remember, that as the Type I error rate (false positive) is reduced, the Type II error rate (false negative) increases! There’s always a need for balance. But that’s a discussion for another day.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

References

1. Fiedler, Janis et al. "Daytime fluctuations of endurance performance in young soccer players: a randomized cross-over trial." *BMC research notes* vol. 15,1 351. 24 Nov. 2022, doi:10.1186/s13104-022-06247-1
2. Top Tip Bio. *The Bonferroni Correction - Clearly Explained*. YouTube, 16 Feb. 2021, <https://youtu.be/HLzS5wPqWR0>. Accessed 23 Dec. 2022.
3. Wikipedia contributors. "Carlo Emilio Bonferroni." *Wikipedia*, 27 Oct. 2022, en.wikipedia.org/wiki/Carlo_Emilio_Bonferroni.
4. Glen, Stephanie. "Holm-Bonferroni Method: Step by Step" From **StatisticsHowTo.com**: Elementary Statistics for the rest of us! <https://www.statisticshowto.com/holm-bonferroni-method/>

Submitted 12/25/2022

IV.14 Inter-Rater Reliability and the Kappa Statistic (Vidal Villela)

Inter-rater reliability

Inter-rater reliability (also known as inter-observer reliability, inter-rater variability, and inter-observer variability) is a metric used in scenarios wherein assessors (or raters) conduct *subjective* judgement on the same variable¹. Inter-rater reliability is an important concern given that distinct individuals assessing data may interpret phenomena differently¹. For example, two clinicians may differ in the degree they grade a pressure ulcer (partly subjectively based on redness and edema), or two independent reviewers assessing publications for risk of bias in a meta-analysis may differ in their assessments. **Intra-rater reliability** is a metric of reliability within a single data collector (i.e., presented the same situation, will an individual interpret this data the same and record the same value each time the variable is presented)¹. In these circumstances, the Kappa statistic provides a useful tool in describing the degree of agreement within a single or between multiple raters.

The Kappa Statistic

Kappa, symbolized by the lower case Greek letter “ κ ”, is a statistic that measures the degree of agreement between a number of raters adjusted for the amount of agreement that would have occurred by chance alone². It is used in scenarios where responses by raters can fall into any number of categories². Its value can range from **-1 to +1**, similar to correlation coefficients¹. Zero in this case represents the amount of agreement represented by chance alone (i.e. no agreement) and 1 represents perfect agreement¹. Not often encountered in practice, -1 theoretically represents perfect disagreement¹. Commonly used iterations of the Kappa statistic are Cohen’s Kappa (used to assess agreement between *two* raters) and Fleiss’ Kappa (used for scenarios with *more than two* raters)².

Interpreting Kappa

Once a number has been generated, if agreement is favored (i.e. $K > 0$), there are several metrics for the ascribed strength of the agreement^{1, 2}. However, note that no universally implemented method exists. A commonly used and accepted metric is the Landis and Koch scale noted below².

Value of Kappa	Strength of Agreement
< 0.00	Poor (worse than chance)
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very Good

Case Example

Suppose a hospital wants to ensure its radiologists are providing reads that are consistent across the department, the Kappa Statistic can be calculated to measure inter-rater reliability across radiologists. In the following scenario, two physicians are independently assessing 100 chest x-rays and determining them to be normal or abnormal, as detailed below.

		Doctor B		
Doctor A	Normal	Abnormal	Total	
Normal	85	7	92	
Abnormal	3	5	8	
Total	88	12	100	

At first glance, one can calculate that the rate of agreement for the aforementioned table is 90% ((85 + 5) / 100), however, this fails to account for the probability of agreeing by chance.

Calculating Kappa^{1, 2}:

**Note that Kappa for >2 observers/raters is not easily calculated by hand and requires statistical software*

$$\text{Kappa} = \frac{P_a - P_c}{1 - P_c}$$

Where P_a is the proportion of categories where there is agreement.

Where P_c is the proportion agreeing by chance, calculated as follows:

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

$$\frac{\frac{\text{Column 1} \times \text{Row 1}}{\text{Number of Observations}} + \frac{\text{Column 2} \times \text{Row 2}}{\text{Number of Observations}}}{\text{Number of Observations}}$$

Proportion where there is agreement = $P_a = \frac{85+5}{100} = 0.90$

Proportion where agreement would be by chance = $P_c = \frac{\frac{(88 \times 92)}{100} + \frac{(12 \times 8)}{100}}{100} = 0.82$

$K = (0.90 - 0.82) / (1 - 0.82) = 0.44$

Applying the aforementioned Landis and Koch scale we can now then determine that, given the **Kappa value of 0.44**, the degree of agreement between both physicians is **moderate**. Below is a different scenario; try to use the aforementioned formulas for Pa, Pc, and Kappa to calculate the Kappa statistic for this scenario.

Doctor B			
Doctor A	Normal	Abnormal	Total
Normal	88	1	89
Abnormal	2	9	11
Total	90	10	100

Here we can see that:

$P_a = 0.97$

$P_c = (80.1 + 1.1) / 100 = 0.812$

$K = (0.97 - 0.812) / (1 - 0.812) = 0.84$

Re-applying our Landis and Koch scale we can determine that here the strength of agreement is **very good**.

Real Application

Below is an abstract from a study published by the Royal College of Radiologists regarding a practical application of the Kappa Statistic³.

Objective: Discrepancy meetings are an important aspect of clinical governance. The Royal College of Radiologists has published advice on how to conduct meetings, suggesting that discrepancies are scored using the scale: 0=no error, 1=minor error, 2=moderate error and 3=major error. We have noticed variation in scores attributed to individual cases by radiologists and have sought to quantify the variation in scoring at our meetings.

Methods: The scores from six discrepancy meetings totalling 161 scored events were collected. The reliability of scoring was measured using Fleiss' kappa, which calculates the degree of agreement in classification.

Results: The number of cases rated at the six meetings ranged from 18 to 31 (mean 27). The number of raters ranged from 11 to 16 (mean 14). Only cases where all the raters scored were included in the analysis. The Fleiss' kappa statistic ranged from 0.12 to 0.20, and mean kappa was 0.17 for the six meetings.

Conclusion: A kappa of 1.0 indicates perfect agreement above chance and 0.0 indicates agreement equal to chance. A rule of thumb is that a kappa ≥ 0.70 indicates adequate interrater agreement. Our mean result of 0.172 shows poor agreement between scorers. This could indicate a problem with the scoring system or may indicate a need for more formal training and agreement in how scores are applied.

Given the importance of accuracy when multiple observers (raters) are interpreting the same phenomenon, the kappa statistic aids in assessing the degree to which agreement is reliably reached.

References:

1. McHugh, Mary L. "Interrater Reliability: The Kappa Statistic." *Biochemia medica* 22.3 (2012): 276–282.
2. Peacock, Janet L., and Phil J. Peacock. *Oxford Handbook of Medical Statistics*, Oxford University Press, Incorporated, 2020.
3. Mucci, B et al. "Interrater variation in scoring radiological discrepancies." *The British journal of radiology* vol. 86,1028 (2013): 20130245. doi:10.1259/bjr.20130245

IV.15 Interim Analyses: When is it justified to prematurely terminate a Clinical Trial? (Navjot Sobti)

Interim analyses are frequently performed in randomized-controlled trials and serve to evaluate the efficacy and/or safety of a new treatment.⁽¹⁾ If, during an interim analysis, the success or futility of a treatment is clearly demonstrated, it may be used to justify the early termination of a clinical trial. Specifically, by virtue of the **Stopping Rule**, if and when the experimental treatment group demonstrates a clear benefit, it is deemed ethical to terminate the clinical trial prematurely. Importantly, interim analyses require “un-blinding” of data⁽²⁾, namely, treatment allocation(s), in order to conduct a comprehensive comparison between the treatment groups. Thus, the interim analysis should ideally be conducted by independent trial statisticians, rather than the primary investigators.

Group Sequential Methods are statistical rules for terminating a study early, when a significant treatment difference is observed during the interim analysis. During a clinical trial, it is not uncommon for study investigators to conduct a series of interim analyses, hence, the term, “sequential.” Typically, two to three interim analyses are deemed sufficient for a clinical trial.⁽⁴⁾ Group sequential methods help to reduce **Type I Error**, which may become inflated during “interim analyses of accumulating data in a clinical trial.”⁽³⁾ Specifically, conducting many interim analyses with a fixed approach of $p < 0.05$ may inflate the false-positive error rate (alpha)⁽⁴⁾, stopping rules help to contain Type I Error by designating low, nominal p values for each interim analysis (**Table 2**). The **“Peto and Haybittle Rule”** and **“O-Brien-Fleming Boundary”** are two examples of commonly used group sequential methods. Both of these methods are “easily implemented,” “adopt stringent criteria (low nominal p-values” and “preserve the intended alpha level and power.”⁽⁵⁾

The Peto and Haybittle Rule,” also known as the “Haybittle-Peto Boundary,” is typically used for randomized controlled trials, as they include both control (e.g., placebo) and experimental treatment groups, in which the “response to treatment is both dichotomous (i.e., success or failure) and immediate.”⁽⁶⁾ This is defined as a **Multiple Testing Procedure**, which is an effective method of “[eliminating] the ethical dilemmas that often accompany clinical trials.”⁽⁶⁾ With the Peto and Haybittle Rule, an interim analysis is performed to evaluate if a statistically significant difference between treatment groups is appreciated, with a p value ≤ 0.001 . If this is the case, the null hypothesis deemed to be true, and early termination of the trial is warranted. When the final analysis is performed, it is evaluated at the normal level (e.g., 0.05) of statistical significance (**Table 2**)⁽⁵⁾. This level of significance is more widely known, and thus, understandable to readers and researchers.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Table 1: Interim stopping levels (p values) for different numbers of planned interim analyses by group sequential design,” from Shulz and Grimes (Lancet, 2005).⁽⁵⁾

Number of planned interim analyses	Interim analysis	Pocock	Peto	O'Brien-Fleming
2	1	0-029	0-001	0-005
	2 (final)	0-029	0-05	0-048
3	1	0-022	0-001	0-0005
	2	0-022	0-001	0-014
	3 (final)	0-022	0-05	0-045
4	1	0-018	0-001	0-0001
	2	0-018	0-001	0-004
	3	0-018	0-001	0-019
	4 (final)	0-018	0-05	0-043
5	1	0-016	0-001	0-00001
	2	0-016	0-001	0-0013
	3	0-016	0-001	0-008
	4	0-016	0-001	0-023
	5 (final)	0-016	0-05	0-041

Overall $\alpha=0.05$.

Table 2: Interim stopping levels (p values) for different numbers of planned interim analyses by group sequential design^{4,25}

Note: the middle column, entitled, “Peto,” represents the Haybittle-Peto Boundaries; the right-sided column, entitled “O’Brien-Fleming,” represents the O’Brien-Fleming Boundary.

In contrast to the Peto and Haybittle Rule, with the **O’Brien-Fleming Boundary**, a different statistical threshold is used at every interim analysis. In general, the O’Brien-Fleming Boundary assigns more stringent statistical thresholds early on, namely, much lower p values. For example, in a trial with three intended interim analyses, the initial p value threshold would be $p < 0.0005$ but would rise to $p < 0.045$ by the final interim analyses (**Table 2**). The O’Brien-Fleming Boundary “[appeals] to many researchers because the stopping criteria are conservative early on, when everyone should be dubious of unstable results, and they successively ease as the results become more stable and reliable.”

Regardless of the stopping rule used for interim analyses, the methodology should be designated prior to the initiation of a clinical trial, and ideally, conducted by independent trial statisticians.

References:

1. Pocock SJ. “When (Not) to Stop a Clinical Trial for Benefit.” JAMA. 2005;294(17):2228–2230. doi:10.1001/jama.294.17.2228
2. Contemp Clin Trials Commun. 2017 Aug 16;7:224-230. doi: 10.1016/j.conctc.2017.08.001. eCollection 2017 Sep.
3. Chen LM, Ibrahim JG, Chu H. Flexible stopping boundaries when changing primary endpoints after unblinded interim analyses. J Biopharm Stat. 2014;24(4):817–833. doi:10.1080/10543406.2014.901341

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

4. "Chapter 20: Multiplicity in Randomised Trials II: Subgroup and Interim Analyses." *Essential Concepts in Clinical Research: Randomised Controlled Trials and Observational Epidemiology*, by Kenneth F. Schulz et al., Elsevier, 2019, pp. 215–225.
5. KF Shulz, DA Grimes. "Multiplicity in randomised trials II: subgroup and interim analyses." *Lancet*, vol. 365 (2005), pp. 1657-1661.
6. O'Brien, Peter C., and Thomas R. Fleming. "A Multiple Testing Procedure for Clinical Trials." *Biometrics*, vol. 35, no. 3, 1979, pp. 549–556. JSTOR, www.jstor.org/stable/2530245.

IV.16 Sensitivity Analysis in Clinical Trials- (Meghan Freed, GSM4)

When reviewing publications for this elective or in your daily work, you may have come across the term “sensitivity analysis” in the statistical analysis section of your manuscript and wondered how this analysis is performed and why it is used. Or, you may have a strong understanding of sensitivity analysis and how this type of analysis can help you in interpreting and determining the “robustness” of the study results. If you are in the former group, this chapter will cover the basics of what a sensitivity analysis is, when you would perform one and the types of analyses that can be performed. If you fall into the latter group, hopefully this chapter will serve as a refresher!

1. What is a sensitivity analysis?

A sensitivity analysis is a statistical method that can evaluate the “robustness” and credibility of a study’s results by changing the primary assumptions, methods, variables, or models that were used to determine the initial results.¹ If the results using new assumptions or methods, etc. DON’T change or are consistent with the initial results, then we can feel more confident about the strength of the primary analysis.^{2,3} If the results DO change, this may lead us to question the initial results.³

2. When would one perform a sensitivity analysis?

There is an element of human error inherent in clinical trials. While we hope that ideal conditions or design of the study will be met, this is not always the case. Patients may miss follow up appointments or fail to fill out questionnaires, leading to missing data. There may be a deviation from protocol or patients that drop out of their specific study group. Sensitivity analyses allow us to test the validity of results when these ideal conditions are not met.³ Examples of scenarios that can arise in clinical trials and create cause for performing a sensitivity analysis include having missing data, protocol deviation or non-adherence to protocol, deciding how to include outlier data, and imbalances in baseline characteristics.² Sensitivity analysis itself is a broad term and does not define a specific type of analysis. For each scenario you encounter, there would be specific methods to choose from to perform this analysis. Below is a table from Thabane et. al that gives examples of these scenarios and what methods you might choose to conduct your analysis.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Table: “Examples of common scenarios for sensitivity analyses in clinical trials”²

(Adapted from Thabane et al.)

Scenario	Sensitivity analysis options
Outliers	<ul style="list-style-type: none"> - Assess outlier by z-score or boxplot - Perform analysis with and without outliers
Non-compliance or protocol violations in RCTs	Perform: <ul style="list-style-type: none"> - intention-to-treat analysis (as primary analysis) - as-treated analysis - per-protocol analysis
Missing data	<ul style="list-style-type: none"> - Analyze only complete cases - Impute the missing data using single or multiple imputation methods and redo the analysis
Definitions of outcomes	<ul style="list-style-type: none"> - Perform analyses on outcomes of different cut-offs or definitions
Clustering or correlation and multi-center trials	<ul style="list-style-type: none"> - Compare the analysis that ignores clustering with one primary method chosen to account for clustering - Compare the analysis that ignores clustering with several methods of accounting for clustering - Perform analysis with and without adjusting for center - Use different methods of adjusting for center
Competing risks in RCTs	<ul style="list-style-type: none"> - Perform a survival analysis for each event separately - Use a proportional sub-distribution hazard model (Fine & Grey approach) - Fit one model by taking into account all the competing risks together
Baseline imbalance	Perform: <ul style="list-style-type: none"> - Analysis with and without adjustment for baseline characteristics - Analysis with different methods of adjusting for baseline imbalance, e.g., Multivariable regression vs. propensity score method
Distributional Assumptions	Perform analyses under different distributional assumptions

	<ul style="list-style-type: none">- Different distributions (e.g., Poisson vs. Negative binomial)- Parametric vs. non-parametric methods- Classical vs. Bayesian methods- Different prior distributions
--	--

3. Example of a sensitivity analysis from the literature when there is missing data

In Voskoboinik et al, researchers conducted a RCT examining the effect of alcohol abstinence on recurrence of atrial fibrillation (afib) and overall afib burden in patients categorized as “regular drinkers” with a history of afib. During the trial, researchers encountered difficulty with patients not attending follow-up visits, failing to fill out or return questionnaires and with secondary outcome measures having missing data.⁴ For their primary endpoints of time to afib recurrence and overall afib burden, there was a low percentage of missing data (0.7% for the abstinence group and 1.4% for the control group).⁴ Due to this low percentage, researchers decided not to employ an alternative statistical method to handle the missing data.⁴ However, there were higher rates of missing data for the secondary endpoints and in this case researchers chose to use a multiple imputations method.⁴ Multiple imputations assumes that the data missing are “missing at random” and can lead to more valid results than other methods of handling missing data.² Otherwise defined, “multiple imputations is a general approach to the problem of missing data that is available in several commonly used statistical packages. It aims to allow for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each of them.”⁵ The results from the multiple imputations method were then compared to a “complete case analysis”, which does not use the missing data.² When results from the imputation method and complete case analysis were compared, researchers noted the results were similar.⁴ Similar results using these two different methods of approaching the missing data gives us more confidence in the validity of the results.

If this type of analysis can help to validate a particular study’s results, why are these not performed in most of the studies we review? Good question. It is argued that this analysis is not performed enough in the medical literature and is actually seen more frequently used in the health economics literature.²

For further reading on sensitivity analyses, the references below are good places to start.

References:

- 1 Porta, M. (Ed.). (2014). *A dictionary of epidemiology*. Oxford university press.
- 2 Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., ... & Debono, V. B. (2013). A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC medical research methodology*, 13(1), 92.
- 3 de Souza, R. J., Eisen, R. B., Perera, S., Bantoto, B., Bawor, M., Dennis, B. B., ... & Thabane, L. (2016). Best (but oft-forgotten) practices: sensitivity analyses in randomized controlled trials. *The American journal of clinical nutrition*, 103(1), 5-17.

- 4 Voskoboinik, A., Kalman, J. M., De Silva, A., Nicholls, T., Costello, B., Nanayakkara, S., ... & Wong, G. (2020). Alcohol Abstinence in Drinkers with Atrial Fibrillation. *New England Journal of Medicine*, 382(1), 20-28.
- 5 Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338, b2393

IV.17 Dealing with Missing Data- what's a person to do? (Daniel Forsman, GSM4)

While working on the neurology service, we had a patient who complained of sporadic syncopal events. The patient had been directly admitted to the hospital for evaluation because their primary care provider was concerned about possible seizures; however, we thought that they simply had orthostatic hypotension and ordered measurement of orthostatics. Unfortunately, later that day, no blood pressures (BP) had been recorded in the electronic medical record (EMR). Although not a research study, this absence of information prevented us from making an accurate diagnosis.

This anecdote is not an anomaly, and missing data is all too common not only in the clinical setting but also in research. When conducting a study, missing data is inevitable. Equipment will break, researchers will make mistakes, and participants will be lost to follow-up. This missing data can drastically impair the conclusions we can derive from a study. This is because most statistical models operate only on complete observations of exposures and outcome variables, which requires researchers to either delete incomplete observations or replace any missing values with estimated values¹. Neither method is perfect, however, and both can introduce bias if used in suboptimal situations. Therefore, it is crucial to understand the nature of any missing data, as the mechanisms that cause data to be missing determine which methods should be used to correct it.

In this chapter, we will discuss the “mechanisms of missing data,” describe how to identify the mechanisms, and conclude with a brief discussion on how researchers deal with missing data.

Types of Missing Data

The framework most commonly used to describe the mechanisms of missing data was devised in 1976 by Donald Rubin². Under this framework, data is classified into one of three different categories based on the relationship between the missing and observed data³. The three categories that comprise this framework are: missing completely at random (MCAR), missing not at random (MNAR), and missing at random (MAR).

Missing Completely at Random: Data is said to be MCAR if the patients who have missing data are a random subset of the complete sample of patients. Additionally, there is no relationship between the missing data and any other values, either observed or missing^{4,5}. Stated more simply, data is classified as MCAR when the reason it is missing is completely random and there is nothing systematic occurring that makes some data more likely to be missing than others⁶. For example, if data is missing because someone dropped a patient's blood sample or because a physician forgot to record a patient's gender, then the missing data would be MCAR. When data is MCAR, no bias is introduced because the set of subjects in the study with no missing data is also a random sample of the population.

Missing Not at Random: When the probability of data being missing is influenced by a missing value, the data is said to be MNAR. An example of this could be a study that is trying to determine the average income for a population. If individuals with higher incomes are less likely to reveal them on a survey than individuals with lower incomes, then the reason data is missing is not random and it would be very difficult to get an accurate picture of the mean income⁴. In this case, we cannot adjust our analyses without strong assumptions and analyses of the data will yield biased results. There is no universal method for handling the missing data properly³, and if you identify that your missing data is MNAR, it is important to perform sensitivity analyses to assess the impact of the missing data⁴. Meghan Freed (chapter XX) provided a well-written summary of sensitivity analyses for those interested.

Missing at Random: MAR lies in between the two extremes and refers to missing data when the reason for the missingness is based only on patient characteristics that we have observed¹. For example, if women are more likely to tell you their weight than men, missing weight values would be MAR. In this situation, there is no relationship between data being missing and unobserved variables, so we can still generate a random subset of the population if we control for the variable that is causing the data to be missing. However, some techniques for handling the missing data, such as simply deleting all data with missing observations, will likely yield biased results.

This topic can be difficult to understand and for the sake of clarity, I have included another example of each mechanism in the excerpt below from Larkins et al.⁶

As an example, let us say we are studying the relationship between blood pressure (BP) and body mass index (BMI) using data from a large national survey and assume that BP is positively correlated with both BMI and anxiety. If on some random visits, the sphygmomanometer was faulty and no BP were recorded, then the missing data on BP might reasonably be assumed to be MCAR. However, some data will be missing because participants declined to have their BP measured, and it might be that people who are obese are more likely to decline BP measurement. In this situation, the average BP of the population would be underestimated using complete case analysis (using only data for participants with observed data for all variables). At the same time, the observed relationship between BP and BMI would still hold true, as the data are MAR. Another possibility is that the most likely group to decline being weighed are those who are both overweight and anxious. If we have no data about participants' anxiety, then part of the data is MNAR. Not only will participants with missing data have a higher average BP, but we will also underestimate their average BP if using BMI only to predict the missing data.

How to Determine the Mechanism of ‘Missingness’ for your Missing Data

Your ability to ascertain the mechanism of missingness depends on your knowledge of the variable being examined in the study⁶. This is especially true when determining whether data is MNAR or not. While researchers could, in theory, follow up on all missing data to allow for comparisons between participants with missing data and those without, this is frequently not possible. Instead, it falls on you to utilize your scientific knowledge and background in medicine to make a reasonable assumption about the data^{5,6}.

Given this uncertainty, a reasonable approach is to begin with the assumption that data are MAR, given that we can usually explain at least some part of the reason data are missing⁶. From there, there are a couple of tests that can be used to determine if your data is MAR or MCAR. First, there is Little’s Test, which many statistical software programs are able to perform. In it, the program tests the null hypothesis that your data are MCAR⁴. Alternatively, you can create a dummy variable that states whether a variable is missing or not. Then, you can run T-tests and chi-square tests between this dummy variable and other variables in your dataset to see if the missingness of the variable is related to other variables⁵. If we did this using our example of MAR data from earlier, we would see that in a chi-square test, the percentage of missing data on weight is higher for men than it is for women.

In the end, these tests are simply data points that you can use. They are **not** absolute. While they may point in one direction or another, it would be unwise to accept these results blindly if they say that the data is MCAR and your knowledge of the missing variable suggests that it is not⁴.

How to Deal with Missing Data

When you are ready to handle your missing data, there are numerous, complex statistical methods that can be used; however, many of these are beyond the scope of this chapter and are far too complex for me to write about with any semblance of eloquence. Instead, I want to give brief descriptions of some of the more common methods, so that when you see them used in studies, you will be able to recognize them and understand if they are being used appropriately.

Complete Case Analysis: In this strategy, all data from participants with at least one missing variable are discarded. For example, if we perform this method on the sample dataset to the right, which is taken from Salgado et al., all of the data from participants 1, 3, 5, 8, and 10 would be deleted. The biggest advantage of using this method is that it is simple, and it is reasonable when the number of data points discarded is relatively small (i.e., less than 5% of the total data; not in this sample's case)^{1,4}. The drawbacks are that it reduces statistical power and that it works under the assumption that the remaining sample is representative of the population as a whole. As we know from before, this will only be true if the missing data is MCAR. If it is used with data that is not, there is a high likelihood that bias will be introduced into your study.

Gender	GLUCOSE	Age
M	?	65
F	120	71
F	99	?
F	140	52
M	88	?
F	85	63
M	170	68
?	153	80
M	115	59
F	103	?

Single Imputation: Imputation is the process of replacing a missing value with a new value. Knowing this, we can logically discern that single imputation is the process of replacing any missing value with a single value (this stands in contrast to multiple imputation). There are many forms of single imputation and common ones include: last observation carried forward (a participant's missing value is replaced by the participant's last observed value), worst observation carried forward (a missing value is replaced by the participant's worst observed value), and simple mean imputation (the missing value is replaced by the mean of that variable)⁴.

Single imputation's benefit is that it does not depend on the data being MCAR, unlike the complete case analysis. One of the cons of this method is that it often results in an underestimation of the variability of a dataset. In addition, it assumes that the value you impute is identical to the missing value. As I am sure you can imagine, this is often an unrealistic assumption, and using this method can often introduce bias into your study⁴. Despite this, Bell et al. found that 27% of all studies assessed used single imputation to replace missing data⁷; therefore, it is likely that you will come across studies that used this method, and when you do, you should be mindful about how it was used.

Multiple Imputation: As mentioned above, single imputation underestimates variability and results in standard errors and P-values that are too small. Multiple imputation solves this problem by introducing uncertainty into the imputed values. There are three main steps to carrying out multiple imputation:

- 1) Imputation: M number of datasets are generated with a different imputed value in each dataset.
- 2) Analysis: Each dataset is analyzed to yield M number of analyses.
- 3) Pooling: All of the M analyses are integrated into one final result.

These are the three basic steps that underlie multiple imputation; however, there are numerous different types of multiple imputation. All of them, though, require that the data is not MNAR to be effective.

Conclusion:

Missing data is ever-present in both the clinic and in research. Furthermore, it will always limit how we can interpret the results of a study. Given its ubiquitous nature, it is important to understand the categories of missing data and how the missingness is best addressed. If we don't, we are susceptible to drawing inappropriate conclusions from our studies.

There is a lot more to this subject, and I did my best to concisely and accurately summarize the topic, but for those who are interested in a more in-depth read, I recommend taking a look at the references below. In particular, Salgado et al. provide a very comprehensive introduction to the topic, and the online statistics blog "The Analysis Factor" has sections that break down these subjects in easy to understand ways. The other references provide more in-depth discussions on missing data.

References:

1. Salgado CM, Azevedo C, Proença H, Vieira SM. Missing data. In: Secondary Analysis of Electronic Health Records. Springer International Publishing; 2016:143-162. doi:10.1007/978-3-319-43742-2_13
2. Rubin DB. INFERENCE AND MISSING DATA. ETS Res Bull Ser. 1975;1975(1):i-19. doi:10.1002/j.2333-8504.1975.tb01053.x
3. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087-1091. doi:10.1016/j.jclinepi.2006.01.014
4. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials-a practical guide with flowcharts. doi:10.1186/s12874-017-0442-1
5. Grace-Martin K. How to Diagnose the Missing Data Mechanism - The Analysis Factor. The Analysis Factor. <https://www.theanalysisfactor.com/missing-data-mechanism/>. Accessed May 21, 2020.
6. Larkins NG, Craig JC, Teixeira-Pinto A. A guide to missing data for the pediatric nephrologist. Pediatr Nephrol. 2019;34(2):223-231. doi:10.1007/s00467-018-3932-4
7. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; A review of the top medical journals. BMC Med Res Methodol. 2014;14(1). doi:10.1186/1471-2288-14-118

Submitted by Daniel Forsman, GSM4, 5/2020

IV.18 Multiple Imputation and Controlled Multiple Imputation: Examples from the OPTION-DM trial (Will Carroll, GSM4)

The two preceding chapters offer an excellent overview of the categorizations of missing data and different imputation methods and are recommended reading. This chapter is an attempt to demystify imputation further by walking through several methods used in the OPTION-DM trial¹ for both MAR (missing at random) and presumed MNAR (missing not at random) data.

1. Background

Briefly, the OPTION-DM trial was designed to assess the efficacy of combined first-line agents for diabetic peripheral neuropathic pain (DPNP) compared to monotherapy in reducing patients' pain levels. Participants rated their pain on the ubiquitous 0-10 pain scale recorded in daily pain diaries. However, due in part to the demanding nature of this trial (more than 20 office visits over 53 weeks), the author's contended with a missing data rate of 15%, (for the primary endpoint)!

2. Step 1 - Multiple imputation assuming MAR in the OPTION-DM trial

First, the authors assumed that the data was missing at random (MAR). To determine the missing values, they used a method known as "nearest neighbors". Specifically, they derived missing scores from a patient's 10 nearest neighbors. This method takes a set of attributes determined by the researcher and determines the "distance" between two participants based on those attributes. A smaller "distance" implies higher similarity, (further processing/weighting of attributes can be performed to generate a "propensity score", but that is beyond the scope of this chapter). By pooling the results of these neighbors, the authors presumably are able to derive feasible pain ratings for an individual's missing pain-scores.

For this trial, neighbors were determined using age, sex, treatment arm, and treatment period. This method is powerful, because it generates plausible values in an intuitive manner (i.e. more similar participants will experience a more similar treatment effect). This, of course, assumes not only that data are missing at random (MAR), but also that participants would continue to behave as if they were still within their specified treatment arm (since 'treatment-arm' is one of the specified attributes).

3. Controlled multiple imputation in the OPTION-DM trial

As stated eloquently in the previous chapter, the true distribution of missing data usually lies somewhere between MCAR and MNAR. When possible, missing-for-cause data should be identified. To assess the impact of potentially meaningful missing data, a method known as "controlled multiple imputation" can be used. This imposes a further assumption on imputed data. Controlled multiple imputation comes in two main flavors, sigma-based and reference based.²

Simply put, sigma-based imputation assumes an offset term (sigma) that describes the difference between the observed and unobserved data. In the case of the OPTION-DM trial, the authors identified potential MNAR data by recording participants' reason for study/treatment discontinuation. They assumed that any participant who halted the study due to poor tolerability or efficacy were

identified as being “for cause”. The main assumption they wanted to explore was that individuals would have *worse outcomes* than predicted by the above multiple imputation method (i.e. higher pain scores). To do this, they described sigma as a range between +0.5 and +2.5. A value from this range was then added to their imputed pain scores, (with a maximum of 10). One can see that if the alternative hypothesis holds water assuming worse outcomes, that this supports the treatment effect observed in their complete-case analysis.

Reference based multiple imputation simply means that assumptions about missing data can be made by referencing other groups of individuals within the trial. This is particularly powerful because it allows one to explore the assumption that an individual with missing data will behave similarly to individuals within a specified trial arm. For example, one could assume that an individual who discontinues the treatment arm of a trial would behave like an individual in the control, or standard-of-care arm, (or any arm for that matter, including the treatment arm)! A very good table and figure are cited from Cro et. al. below, as they illustrate different reference-based multiple imputation methods extremely well.

Conclusion

Missing outcome data is a common problem in RCTs, and multiple imputation is an increasingly utilized method to perform sensitivity analysis, and even to occasionally perform primary analysis³. Therefore, it is important to have a basic understanding of multiple imputation and controlled imputation. When determining the validity of imputation, one should be aware of the assumptions that researchers made and look for studies where sensitivity analysis is performed using multiple assumptions, and that these are explicitly stated somewhere in the primary manuscript or appendix. A complete set of recommendations is provided by Tan et al, 2021⁴, under ‘Future Recommendations’, and we can apply these to the OPTION-DM trial.

For the OPTION-DM trial, they explicitly stated their model and method of multiple imputation (nearest neighbors), and reported the outcomes associated with each assumption. For controlled multiple imputation they explicitly describe the scenario necessitating controlled MI (as above), their rationale, and results of controlled MI analysis. This level of transparency, and the conservative assumption used for controlled MI, help put our minds to rest that the conclusion of the trial (that combination therapy is superior to monotherapy in reducing DPNP), is likely, despite a large proportion of missing data.

Figure and chart describing reference-based MI methods²

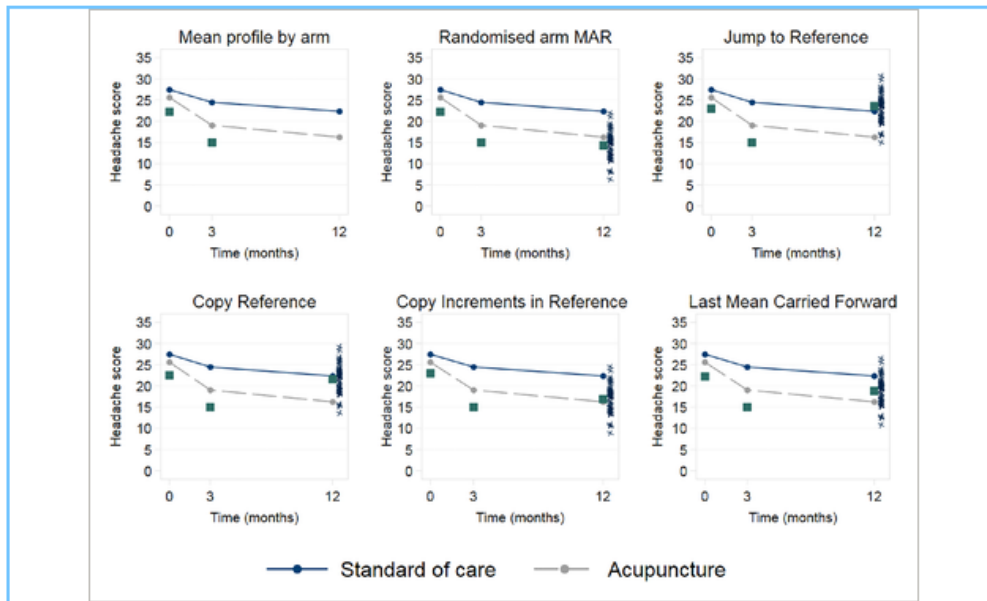


FIGURE 4

[Open in figure viewer](#) | [PowerPoint](#)

Example reference based imputation models for the acupuncture trial. The squares at time 0 and 3 are observed values for a participant in the acupuncture arm who withdrew after the 3 month visit. The black arrows represent the imputation distributions. The squares at time 12 are the mean of the imputed values for that participant in the given reference based scenario. The crosses at time 12 represent the individual imputed values around that mean across 50 multiply imputed datasets for the withdrawing active participant. The reference arm is the standard care arm. This is not an exhaustive display of the MNAR options possible within the reference-based framework [Colour figure can be viewed at wileyonlinelibrary.com]

Method	Description
Randomized-arm MAR	Impute assuming patients follow the behavior of their randomized arm. The joint distribution of patients' pre- and post-deviation outcome data is MVN with mean and covariance matrix from their randomized arm.
Jump to reference (J2R)	Impute assuming patient behavior jumps to that of a specified reference arm. The joint distribution is MVN with mean vector from the patients' randomized arm up to their last observation time, post-deviation the mean vector follows that observed for a reference group (typically control). The covariance matches the randomized arm for pre-deviation measurements and the reference arm for the conditional components of post- given pre-deviation measurements.
Last mean carried forward (LMCF)	Impute assuming patient behavior remains at the mean level for their randomized arm at their last observed time point. The joint distribution is MVN with mean vector from the patients randomized arm up to their last observation time, post-deviation the means are set equal to the marginal mean for the

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Method	Description
	patients randomized arm at their last observed time. The covariance matrix remains as that for their randomized treatment arm.
Copy increments in reference (CIR)	Impute assuming patient behavior follows the mean increments observed in a specified reference arm. The joint distribution is MVN with mean vector from the patients randomized arm up to their last observed time, post-deviation the patients' mean increments follow those from a reference arm. The covariance is the same as in J2R. Appropriate when we wish to assume that post-deviation the disease resumes the course observed in the reference arm.
Copy reference (CR)	Impute assuming patients follow the behavior of a specified reference arm for the duration of the trial. The joint distribution of patients' pre- and post-deviation outcome data is MVN with mean and covariance matrix from a reference arm regardless of deviation time.

References:

¹ Tesfaye, S., Sloan, G., Petrie, J., White, D., Bradburn, M., Julious, S., ... & Selvarajah, D. (2022). *The Lancet*, 400(10353), 680-690. [Comparison of amitriptyline supplemented with pregabalin, pregabalin supplemented with amitriptyline, and duloxetine supplemented with pregabalin for the treatment of diabetic peripheral neuropathic pain \(OPTION-DM\): a multicentre, double-blind, randomised crossover trial - PMC \(nih.gov\)](#)

² Cro, S., Morris, T. P., Kenward, M. G., & Carpenter, J. R. (2020). Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: a practical guide. *Statistics in medicine*, 39(21), 2815-2842.

³ Bell, M. L., Fiero, M., Horton, N. J., & Hsu, C. H. (2014). Handling missing data in RCTs; a review of the top medical journals. *BMC medical research methodology*, 14(1), 1-8.

⁴ Tan, P. T., Cro, S., Van Vogt, E., Szigeti, M., & Cornelius, V. R. (2021). A review of the use of controlled multiple imputation in randomised controlled trials with missing outcome data. *BMC Medical Research Methodology*, 21(1), 1-17.

Submitted 1-8-2024

IV.19 Reporting and Interpreting Economic Analysis (Ashley Baronner)

Economic analyses can be difficult to report and interpret and can also be very demanding to conduct. The general goal of most economic analyses is to encourage clinicians and readers to think more laterally about cost. Rather than thinking linearly about the defined cost of a treatment, economic analyses look at “opportunity cost” and resource utilization. In reading and interpreting an economic analysis, it is important to pay attention to validity, importance, and applicability.

1. Validity

- a. Are well-defined courses of action compared?
 - i. For example, in a study of IV iron comparing this method to oral iron supplementation, liquid supplementation, and blood transfusion rather than focusing solely on the study intervention.
- b. Does it provide a specified view from which costs and consequences are being assessed?
 - i. Is the cost viewed from the experience of the individual patient, hospital, government, or entire population?
- c. Does it cite comprehensive evidence on the efficacy of alternatives?
- d. Does it identify all the costs and consequences and select credible and accurate measures of them?
 - i. This includes direct costs (hospitalization, medication cost) and indirect cost (time lost from work).
- e. Was the type of analysis appropriate for the question posed?
 - i. One option is cost effectiveness analysis, which cannot compare different types of health outcomes (see more below).
 - ii. Another option is a “cost-benefit” analysis, although this possesses the challenge of assigning monetary value to life itself.
 - iii. Another option is a “cost-utility” analysis. This involves framing outcomes in terms of desirability of a certain outcome such as morbidities associate with and without treatment. Utility plus time will generate QALYs (quality adjusted life years- generic measure of disease burden, including both the quality and the quantity of life lived)
 1. Example of QALY: 1 year of perfect health is equivalent to 2 years post disease in a state of decreased utility, 0.5.

2. Importance

- a. Are the resulting costs or cost/ unit of health gained clinically significant?

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

- i. Cost-minimization analysis takes into account if the cost difference is large enough to warrant changing the standard of practice.
 - b. Did the results of this economic analysis change with sensible changes to costs and effectiveness?
 - i. Cost-effectiveness considers whether the difference in effectiveness is sufficient to spend the difference between costs.
3. Applicability
 - a. Do the costs in the economic analysis apply in our setting?
 - b. Are the treatments likely to be effective in our setting?

Achieving economic appraisal

The core components of economic appraisal include cost-effectiveness analysis, cost-benefit analysis and cost utility analysis.

Cost-effectiveness (C/E) is typically expressed in terms of monetary cost per case of a disease. For example, different programs to control hyperlipidemia could be compared in terms of dollars saved per degree of reduction in cholesterol. However, disparate outcomes such as cost of reduction in cholesterol verses cost of avoided hospitalizations for MIs cannot be compared as the denominators for C/E will be different. In addition, cost-effectiveness cannot be used to compare multiple clinical effects such as morbidity and mortality.

Cost-benefit analysis (CBA) determines the net social benefit of the program (NSB). If the NSB is greater than zero, it should be implemented. If it is less than zero, it should not. Cost-benefit analysis is useful because unlike cost-effectiveness, disparate effects (such as morbidity and mortality) can be compared. Net resource utilization should not be used alone without other measure to determine the value of consequences and the value of improved health itself.

Cost-utility analysis (CUA) incorporates QALYs (quality adjusted life years) gained (see example above). CUA is often considered to be more compatible with the way health care providers make decisions, as quality of life is factored into health outcomes. This is especially useful when the intervention being evaluated impacts both morbidity and mortality and a common unit of outcome combining both is desired. In addition, CUA from different interventions can be compared. In order to use CUA, effectiveness data for health outcomes must be available.

Ethics and broader implications

Economic analyses reveal important trade-offs when considering alternative interventions with the goal being as much health improvement as possible with the available resources. These analyses help to define the potential health benefits lost when the best alternative is not selected. The CDC commonly uses economic analyses in their deliberations regarding topics such as vaccines. However, Medicare does not use cost-effectiveness in determining whether or not to cover new therapies. The Affordable Care Act forbids use of QALYs, possibly due to underlying mistrust of underlying methods or desire to downplay resource scarcity in health care. Ethical considerations by the US Public Health Service recommend the use of cost-effectiveness analysis with attention to “societal perspective” reflecting population health as well as a health care sector perspective. Other ethical considerations such as distributive concerns and non-health related effects of interventions should be factored in as well. These factors are all relevant in determining how to spend society’s limited resources on health care. Accepting trade-offs is an essential, but challenging, aspect of cost effectiveness analysis.

Example:

Let’s take a look at how one might interpret the economic analysis of male HPV vaccination in the United States. Cost effectiveness of vaccines is influenced by vaccine efficacy, durability, severity of disease burden, vaccine price, and delivery-program costs. The HPV vaccine costs \$109 per dose. Full vaccination which includes more than 30 doses against 16 diseases costs approximately \$1,450 in males and \$1800 in females. In expanding the HPV vaccine to males, one must take into account if covering the cost of the vaccine verses alternative use of these dollars such as improving vaccine uptake in girls.

In a study of the cost effectiveness of HPV vaccinations in the United States, male vaccination cost effectiveness depended on vaccine coverage of females. When all HPV associated outcomes in the analysis were considered, the incremental cost per quality-adjusted life years (QALY) gained by adding male vaccination to a female only vaccination strategy was \$23,600 in the lower female coverage scenario and \$184,300 in the higher female coverage scenario. Including male vaccination appeared less favorable in terms of cost effectiveness when compared to a strategy of increasing female vaccination coverage. In terms of validity, this study focuses on the quadrivalent HPV vaccine, which was the only vaccine of its kind at the time of the study. The cost analysis is appropriate, but the study does assume 100% vaccination success rate. There are no alternative vaccines to investigate, but the strategy of including males verses expanding female coverage was investigated. This study looks at cost from a population based standpoint with the goal being to reduce the overall burden of HPV. This study includes cost-effectiveness analysis and cost utility analysis with QALYs. However, QALYs with respect to HPV-related health outcomes provide a source of uncertainty, which is subject to change with more data regarding non-cervical cancers associated with HPV. Cost utility analysis is appropriate for this type of intervention in which investigators want to factor quality of life into health outcomes. Cost benefit analysis was not performed. The implications are important, and clinically applicable to this patient population.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Resources:

- 1 Chesson, Harrell W., et al. "The cost-effectiveness of male HPV vaccination in the United States." *Vaccine* 29.46 (2011): 8443-8450.
- 2 Guyatt, Gordon H., et al. "Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care." *Jama* 284.10 (2000): 1290-1296.
- 3 Henry, David. "Economic analysis as an aid to subsidisation decisions." *PharmacoEconomics* 1.1 (1992): 54-67.
- 4 Kim, Jane J. "The role of cost-effectiveness in US vaccination policy." *New England Journal of Medicine* 365.19 (2011): 1760-1761.
- 5 Kerridge, Ian, Michael Lowe, and David Henry. "Ethics and evidence based medicine." *Bmj* 316.7138 (1998): 1151-1153.
- 6 Neumann, Peter J., and Gillian D. Sanders. "Cost-effectiveness analysis 2.0." *N Engl J Med* 376.3 (2017): 203-5.
- 7 Torrance, George W. "Measurement of health state utilities for economic appraisal: a review." *Journal of health economics* 5.1 (1986): 1-30.

February 2019

IV.20 The Internists Guide to Choosing the Correct Statistical Test (J.D. Nuschke III)

Imagine you are in the following scenario: You have a clinical question that you believe could make for an interesting paper. You have access to a database with large amount of information on the subject of interest.

Do you know how to answer your question with the data you have?

If the answer is no, then you are reading the correct outline entitled “The Internists Guide to Choosing the Correct Statistical Test”. If the answer is yes, then I am impressed, and you can move on.

Over the following pages, we will examine how to choose the correct statistical test for the data you are analyzing through a series of questions. It may be helpful to print this page, and circle your answers to each question. At the end there will be a Graph that you can utilize (based on the answers to the questions below) to choose the most appropriate statistical test.

Please note, this guide will make the following assumptions-

- Reader understands Descriptive Statistics (mean/median/mode/standard deviation).
- Reader understands the role of inferential analysis (we use statistical tests to determine if patterns are due to chance vs intervention effect).

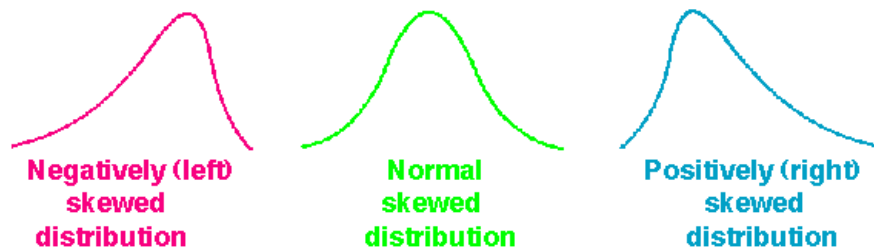
- First question- How many Dependent and Independent variables do you have?
 - Dependent Variable- What you measure in the experiment and what is affected
 - Independent variable- the variable that is changed or controlled
 - Example: You are interested in if 5 mg of amlodipine controls blood pressure better than placebo.
 - Dependent Variable- Blood Pressure
 - Independent Variable- Dose of amlodipine
- Second Question - What Type of variables do you have? Note: Answer for both dependent and independent variables
 - Variable Type-
 - **Categorical**- Made of categories
 - Binary –Self-explanatory- EX: 0 and 1
 - Nominal- Think “NAME inal”- No value to category. EX: Hair Color
 - A variable may be coded, then, with a number. This number has no value. EX blonde 1, brunette 2, red hair 3.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

- Ordinal- Think “Order inal”
 - The order of the values are placed in logical order, but without known increments
 - EX: Income: 1-low, 2- medium, 3- high
 - The above income categories could be based on a predetermined amount (EX 10, 50, 100 dollars). Therefor a High income will not mean 3 times more money than low income.
 - ****TIP****The above concept, while it may seem simple, is a key distinction between CATEGORICAL and CONTINUOUS variables and will be explored in the continuous variable section
- **Continuous**- Any score of value within a measurement scale AND the differences have meaning (see directly above)
 - Interval- Variable that can be ordered and the distance between variables is meaningful
 - EX Temperature:98, 99, 100 degrees etc
 - Ratio- Variable that can be ordered, distance between variables is meaningful, HAS A 0 POINT
 - The Zero point allows the ratio of the score to make sense
 - EX Age: 0 years, 1 years, 2 years.... 20 years.
 - You could then say, a 10-year-old is half as old as a 20-year-old. You’ve just created a ratio
- Third Question- What are you measuring?
 - Categorical variables
 - Mean
 - Median
 - Proportions
 - Continuous variables
- Fourth Question- What does your data look like?
 - Normal Distribution
 - EX: Your data would fit well under a bell curve
 - Non- Normal
 - Skewed

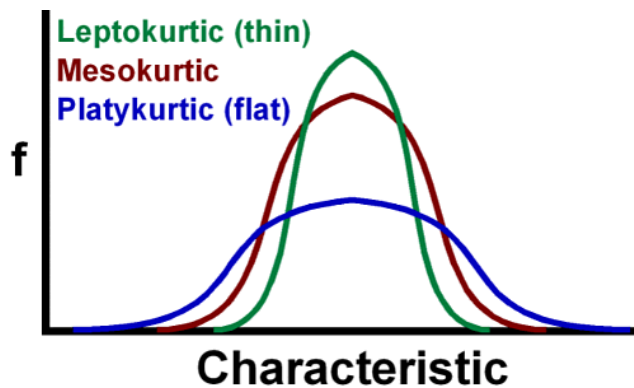
Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

- Negative skew- most scores at higher end of possible scores
- Positive skew- most scores are at lower end of possible scores
- Kurtosis
 - Leptokurtic-The values exhibit a peak in the middle
 - Platykurtic- The values exhibit a broad range of similar values
- Graphical Depiction



Reference:

- 1 <https://researchbasics.education.uconn.edu/normal-distribution/>



Reference:

- 2 <https://www.quora.com/How-can-I-understand-different-types-of-kurtosis>

After answering all 4 of the prerequisite questions, you should have the tools to choose the correct statistical test using the guide below

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Number of Dependent Variables	Number of Independent Variables	Type of Dependent Variable(s)	Type of Independent Variable(s)	Measure	Test(s)
1	0 (1 population)	continuous normal	not applicable (none)	mean	one-sample t-test
		continuous non-normal		median	one-sample median
		categorical		proportions	Chi Square goodness-of-fit, binomial test
	1 (2 independent populations)	normal	2 categories	mean	2 independent sample t-test
		non-normal		medians	Mann Whitney, Wilcoxon rank sum test
		categorical		proportions	Chi square test Fisher's Exact test
	0 (1 population measured twice) or 1 (2 matched populations)	normal	not applicable/ categorical	means	paired t-test
		non-normal		medians	Wilcoxon signed ranks test
		categorical		proportions	McNemar, Chi-square test
	1 (3 or more populations)	normal	categorical	means	one-way ANOVA
		non-normal		medians	Kruskal Wallis
		categorical		proportions	Chi square test
	2 or more (e.g., 2-way ANOVA)	normal	categorical	means	Factorial ANOVA
		non-normal		medians	Friedman test
		categorical		proportions	log-linear, logistic regression
	0 (1 population measured 3 or more times)	normal	not applicable	means	Repeated measures ANOVA
	1	normal	continuous	correlation	
		non-normal		simple linear regression	
		categorical	categorical or continuous	non-parametric correlation	logistic regression
	2 or more	normal	continuous	continuous	discriminant analysis
continuous				multiple linear regression	
continuous				logistic regression	
non-normal		mixed categorical and continuous	Analysis of Covariance General Linear Models (regression)		
			continuous	logistic regression	
			continuous	logistic regression	
2	2 or more	normal	categorical	MANOVA	
2 or more	2 or more	normal	continuous	multivariate multiple linear regression	
2 sets of 2 or more	0	normal	not applicable	canonical correlation	
2 or more	0	normal	not applicable	factor analysis	

Source: James D. Leeper, Ph.D. (University of Alabama)

So now you've

- 1) thought of a clinical question
- 2) defined your variables
- 3) defined your measurements
- 4) defined the distribution of your data
- 5) utilized the above graph to choose the correct statistical test

Congratulations! You are now ready to analyze your data!

Attached below, for your convenience, are examples of the most commonly used statistical tests.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Example of commonly used tests

Chi Square Test- Tests the strength of association between two Categorical Variables

- EX: Is being on amlodipine (independent variable with 2 categorical options- on amlodipine or not on amlodipine) associated with normotensive blood pressure (dependent categorical variable with 2 categorical options- normotensive or not normotensive)

2 independent sample T test- Test for the difference between two independent variables

- EX: Do patients on amlodipine (independent variable with 2 categorical options- on amlodipine or not on amlodipine) have lower average systolic blood pressure (dependent variable, mean measurement, assuming normal distribution) than patients on placebo.

Paired T test- Tests for the difference between two related variables

- EX: Do patients have lower average systolic blood pressure (dependent variable, mean measurement, assuming normal distribution) on amlodipine than when the same patients are not on amlodipine (independent paired variable as SAME PATIENTS have 2 categorical options- on amlodipine or not on amlodipine)

ANOVA- Tests the hypothesis that mean values of dependent variable are different between 2 or more group means AFTER any other invariance in outcome variable is accounted for

- EX: Do patients on 10 mg, 5 mg and 0 mg of amlodipine (independent variable with 3 categorical options- 10, 5 and 0 mg) have a difference in mean systolic blood pressure (dependent variable, mean measurement, assuming normal distribution)

Linear Regression- Test that sees whether a variation in the independent variable causes variation in the dependent variable. This allows you to estimate the correspondence of one unmeasured variable to a measured variable.

- EX: Does 0,5,10,15,20mg of amlodipine (independent variable with multiple options) create a dose dependent response in mean systolic blood pressure (dependent variable, mean measurement, assuming normal distribution)?
- With this information, you be able to choose the correct dose based on how a patient is responding to current dose (all other factors being equal).

The ideas of more advanced statistical tests (IE-logistic regression, friedman tests, you-name-it test) are all built around the foundation we laid today!

Armed with this information, you should be able to tackle the basic statistical tests we use in clinical medicine, sound smart and help your patients through the science of evidence based medicine.

References (Images/Charts Referenced below image):

- 1 "Types of Statistical Tests." CYFAR, University of Minnesota, 2019, cyfar.org/types-statistical-tests.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

- 2 McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.

April 2019

Section V. Finding, Appraising, and Applying Evidence

V.1 Health literacy and numeracy – an essential feature of evidence based medicine- Caroline Lombardo

Limited health literacy and numeracy can present challenges within the healthcare encounter, which in turn affect outcomes of morbidity and mortality. Evidence-based medicine (EBM) and its application to individual patients requires engagement on part of the patient and their participation in shared-decision making. This process of bringing scientific knowledge from bench to bedside (or into the office!) has several barriers, one of which can be a patient's limited capacity to understand and implement health information.

The concept of "health literacy" broadly encompasses the abilities of patients to work with knowledge and information needed to maintain and improve their health, taking into account both individual and system contexts (1). Additionally, health numeracy can impact a patient's ability to make healthcare decisions based on numerical information, which is often an important component of EBM.

The general practice of explaining and disseminating health information in both verbal and written formats at a 6th to 8th grade reading level is recommended by the American Medical Association (AMA) and National Institutes of Health (NIH) respectively (2). While these recommendations are widely known by physicians, the majority of written patient education materials in high-impact medical journals continues to be above these recommended readability grades (one analysis identified a mean readability grade range of 11.2 to 13.8 (3)).

Currently, some health-decision aids are available to use within the patient encounter to illustrate concepts of risk using visual displays. These decision-aids were constructed utilizing results from studies that suggest combined numerical and graphical displays of information can aid in the communication of risk (4, 5).

One challenge to the practice of EBM is the patient who has limited health literacy and presents with medical information that is wrong or misinterpreted. Oftentimes, this information comes from the internet, where patients have access to an unlimited and unfiltered source of health information. Many people however don't have the ability to discriminate whether this information is true or applicable to their own health, further highlighting the need for a strong skillset in health communication when practicing EBM.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

As physicians, there are various approaches we can take to improve the communicability of EBM concepts when working with patients that have limited health literacy and numeracy. Currently, some electronic medical record (EMR) systems integrate decision-aids to assist with the depiction of risk and other potentially complicated numerical concepts. Patient education materials written at an appropriate readability level are also important to utilize in order to adequately inform the patient with limited health literacy. The next step in the process of improving health literacy should include efforts to bring forward conversations around complicated health topics to a broader lay audience and to ensure that important concepts in EBM are included such as critical appraisal of study methodology, risk, sensitivity, and specificity.

Perhaps the most important additional considerations include 1) the time clinicians take (or are allowed to take) to explain study findings or how they impact the recommended therapies, 2) the inclination to embrace shared decision making, and 3) the preparedness of the practitioner to practice and to teach EBM to colleagues, patients and other learners. Each of the points referred to in the foregoing paragraphs could spawn a detailed exploration which this writer hopes to address as she moves forward.

References:

1. Liu C, Wang D, Liu C, et al. What is the meaning of health literacy? A systematic review and qualitative synthesis. *Family Medicine and Community Health*. 2020;8(2):e000351. doi:<https://doi.org/10.1136/fmch-2020-00035>
2. Weiss B. Health Literacy: A Manual for Clinicians Health Literacy a Manual for Clinicians Part of an Educational Program about Health Literacy. <http://lib.ncfh.org/pdfs/6617.pdf>
3. Rooney MK, Santiago G, Perni S, et al. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. *Journal of Patient Experience*. 2021;8:237437352199884. doi:<https://doi.org/10.1177/2374373521998847>
4. Hamstra DA, Johnson SB, Daignault S, et al. The Impact of Numeracy on Verbatim Knowledge of the Longitudinal Risk for Prostate Cancer Recurrence following Radiation Therapy. *Medical Decision Making*. 2014;35(1):27-36. doi:<https://doi.org/10.1177/0272989x14551639>
5. CDC. Health Literacy Research and Best Practices. Centers for Disease Control and Prevention. Published March 18, 2022. <https://www.cdc.gov/healthliteracy/researchevaluate/numeracy.html>

Submitted 6-28-2023

V.2 Obtaining High-Quality Studies and Clinical Guidelines: A User-Friendly Overview (Alexander Kettering, GSM4)

Whether within the confines of daily clinical practice, basic science research, or even a general interest in research, it can be difficult to have a streamlined approach to finding high-quality, valid studies and guides or to clinical practice guidelines. The goal of this section will be to introduce readers to a straightforward approach to three reliable databases that provide comprehensive, but unique evidence-based resources and guidelines. More specifically, the entry will focus on UpToDate, DynaMed Plus, and ACCESSSS of McMaster University.

General Approach and Introduction to Sourcing Reliable Data

When trying to decide on the best database to use, one must first decide on the nature of the question being asked, or the primary purpose of a particular search; what is that question that one is trying to answer? What kind of study could help answer that question in the most effective way? Due to an inherent ability to control for factors such as confounding, as well as a marked ability to assess for causality between variables, randomized control trials (RCTs) are optimal. Other types of studies, such as cohort studies, can be helpful in certain circumstances, such as when one wants to add additional value or insight into a question, but they do not establish causality. Therefore, whether the goal is to simply find a quick reference while on the clinical wards to answer a question about patient care in real-time, or if the goal is to perform a more in-depth analysis of a particular research question, the general consensus is that guidelines and data based in randomized control trials is generally preferred. All three of the databases outlined in this entry do a reliable job of providing references based in randomized control trials.

The next step is to answer just the question proposed above: what is the goal of one's search? In what context is the search being performed? The answer to this question will help dictate the optimal database. If one is perhaps trying to answer a question quickly amidst the ebb and flow of inpatient rounds, DynaMed Plus might be the best resource to start with. If one has a little bit more time to sit down and obtain a more in-depth overview of a particular topic, but the goal is not necessarily to dive straight into original literature, then UpToDate may be the best database in that instance. These two databases provide real-time material that is updated by editors on a regular basis. They provide clinical guidance without interrupting the day's workflow to a great extent, and simultaneously provide citation links to original studies. The final database that is of note is ACCESSSS (<https://www.accessss.org/>), which is a sort of "master database" created by McMaster University. ACCESSSS provides extremely efficient, well-organized access to pre-reviewed studies that meet a very specific set of criteria to assure quality and validity. In the following section, provided will be a quick-reference list of the pros and cons of each of the individual databases introduced above.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

The Pros and the Cons

UpToDate (www.UpToDate.com)

Pros	Cons
Just as it sounds, it is regularly updated and tends to have the latest and clinically important information and links to journal articles.	Very heavy in volume, more in the “prose” format.
Many institutions regard it as the Gold Standard, go-to resource for evidence-based care while managing patients either on the wards or in an ambulatory setting.	Can often feel as if there is a lot to sift through in order to find answers – may lead to compromise of efficiency.
Contains extremely clear tables, flowcharts, and graphs that give evidence-based approaches to clinical decision making.	
A very large database, very little that is not present with fleshed out references that appear to be much more extensive than some other resources.	
Overall, a high-quality “first pass” for data gathering and clinical guidance. If you need a reliable place to start, cannot go wrong with UpToDate.	

User Interface and Sample Search:

The screenshot shows the UpToDate website interface. At the top, there is a search bar with the text 'atrial fibrillation aspirin' and a search button. To the right of the search bar, there are user options: 'Alexander Kettering', 'CME 193.0', and 'Log Out'. Below the search bar, there is a navigation menu with 'Contents', 'Calculators', and 'Drug Interactions'. A 'Back to Search' button is also visible. The main content area displays the title 'Atrial fibrillation: Anticoagulant therapy to prevent thromboembolism' along with author and editor information. The article text begins with an introduction discussing the development and subsequent embolization of atrial thrombi.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

DynaMed Plus (www.dynamed.com)

Pros	Cons
By far the most efficient resource for real-time use on the wards.	Often, tends not to be updated as regularly as UpToDate, and often the references are not quite as recent as other resources.
Bullet-point format makes things much more efficient for the practitioner.	At times, provides less detail than other resources, and requires that users reference other resources as well to completely answer certain clinical questions.
Includes succinct examples of characteristics of a particular pathology or presentation, and how said characteristics might present in an HPI, Review of Systems, or Physical Exam. This allows providers to easily reference this database in real-time, even in the midst of a patient visit if necessary!	
References are in a unique format, in that the authors and editors of DynaMed list references with direct links to original papers, and then also give level of evidence based on a scoring system.	

User Interface and Sample Search:

The screenshot displays the DynaMed Plus interface. At the top, there is a search bar with the query "atrial fibrillation aspirin". Below the search bar, the main heading is "Thromboembolic Prophylaxis in Atrial Fibrillation". The page is organized into a sidebar with a "TOPIC" tab selected, and a main content area. The sidebar lists various sections like "Overview and Recommendations", "Background", "Evaluation", "Management", etc. The main content area shows the "Overview and Recommendations" section, which includes a "Background" subsection with bullet points discussing the complications of atrial fibrillation and the use of antithrombotic therapy. On the right side, there is a panel for "TOPIC EDITOR" and "RECOMMENDATIONS EDITOR" with names and credentials. At the bottom right, there is a "Feedback" button.

Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine

ACCESSSS (www.accessss.org)

Pros	Cons
This is the one for a deep-dive; it provides both breadth and depth in one's search.	To some degree, relying on the website's protocols for study localizations – it is possible to miss particular studies that you might have found if self-managed in a database such as PubMed. Less control over granularity of the search, though more user options have been added.
Useful to create a "master list" of original sources that span the entire spectrum of study-types.	
Organizes by Clinical Texts, Guidelines, Systematic Reviews, and Original Studies.	
Particularly nice because it does a lot of the work for you – rather than forcing you to insert special search parameters in a database such as Cochrane or PubMed, this includes references from all of the main resource sites, and pre-vets the resources based on quality, citations, validity, etc..	
In other words, one can nearly certainly rest assured that resources obtained from ACCESSSS will be of the highest quality.	
The database will send users regular emails with updated search results, and individual searches can be saved.	

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

User Interface and Sample Search:

The screenshot displays the ACCESSSS search interface. At the top, the logo 'ACCESSSS SMART SEARCH BEST EVIDENCE FOR HEALTH CARE' is on the left, and navigation links 'Search', 'Articles', 'Dashboard', 'About', 'External Links', 'Help', and 'My Account' are on the right. The search bar contains the text 'atrial fibrillation aspirin'. Below the search bar, there are filters for 'PLUS Database: MD', 'Selected Library: None', 'Your Search History', and 'Advanced Options'. The results are categorized into four main sections: 'Summary Clinical Texts' (DynaMed: 50 Items, Best Practice: 35 Items, EBM Guidelines: 22 Items), 'Systematic Guidelines' (Guidelines in McMaster PLUS: 0 Items), 'Systematic Reviews' (ACP Journal Club: 4 Items, McMaster PLUS: 35 Items), and 'Original Studies' (ACP Journal Club: 13 Items, McMaster PLUS: 49 Items). A message states: 'UpToDate's search results would qualify to be listed here, but they don't allow being incorporated into search engines. Click here to be transferred to UpToDate.' At the bottom, there are buttons for 'Revise this search', 'Start a new search', 'Load a search', and 'Save this search'. The McMaster University and McMaster PLUS logos are visible in the bottom right corner.

V.3 Levels of Evidence and Recommendations- USPSTF and AAFP- (Aditya Kulkarni)

When searching for evidence-based medicine practices among the various clinical guidelines that exist, it can often be difficult to determine the accuracy of recommendation. Describing the strength of a recommendation is an important part of communicating its importance to providers and patients. Fortunately, there are a few different grading systems for ranking recommendations. One commonly used system is the A-I grading system used by the USPSTF. Another system, proposed by the American Academy of Family Practice, is the Strength of Recommendation Taxonomy (SORT). In this chapter brief overviews of both are provided.

I.USPSTF- US Preventive Services Task Force

Grade	Definition	Suggestions for Practice	Example
A	The USPSTF recommends the service. There is high certainty that the net benefit is substantial	Offer or provide this service	1) Colorectal CA screening for adults aged 50-75 years. 2) High blood pressure screening in adults age 18 years or older.
B	The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate or substantial	Offer or provide this service	1) Screening for abnormal blood glucose as part of cardiovascular risk assessment in adults aged 40-70 years who are overweight or obese. 2) Biennial screening mammograms for women aged 50-74 years
C	The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgement and patient preferences. There is at least moderate certainty that the net benefit is small	Offer or provide this service for selected patients depending on individual circumstances	1) Prostate Cancer screening in men aged 55-69 years 2) Biennial screening mammograms for women aged 40 to 49 years.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

D	The USPSTF recommends <i>against</i> the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits	Discourage the use of this service	1) Testicular cancer screening in adolescents and adult males 2) Screening for COPD in adults using spirometry
I	The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined	Read the clinical considerations section of USPSTF Recommendation Statement. If the service is offered, explain to patients the uncertainty about the balance of benefits and harms	1) Screening adults for glaucoma 2) Whole-body skin examination by a PCP or patient skin self-examination for the early detection of cutaneous melanoma, basal cell carcinoma, or squamous cell skin cancer in the adult general population

II. Strength of Recommendation Taxonomy (SORT)

As of 2004, the AAFP has developed a grading scale with the goal of allowing readers of family medicine and primary care journals to have one scale that addresses the quality, quantity, and consistency of evidence.

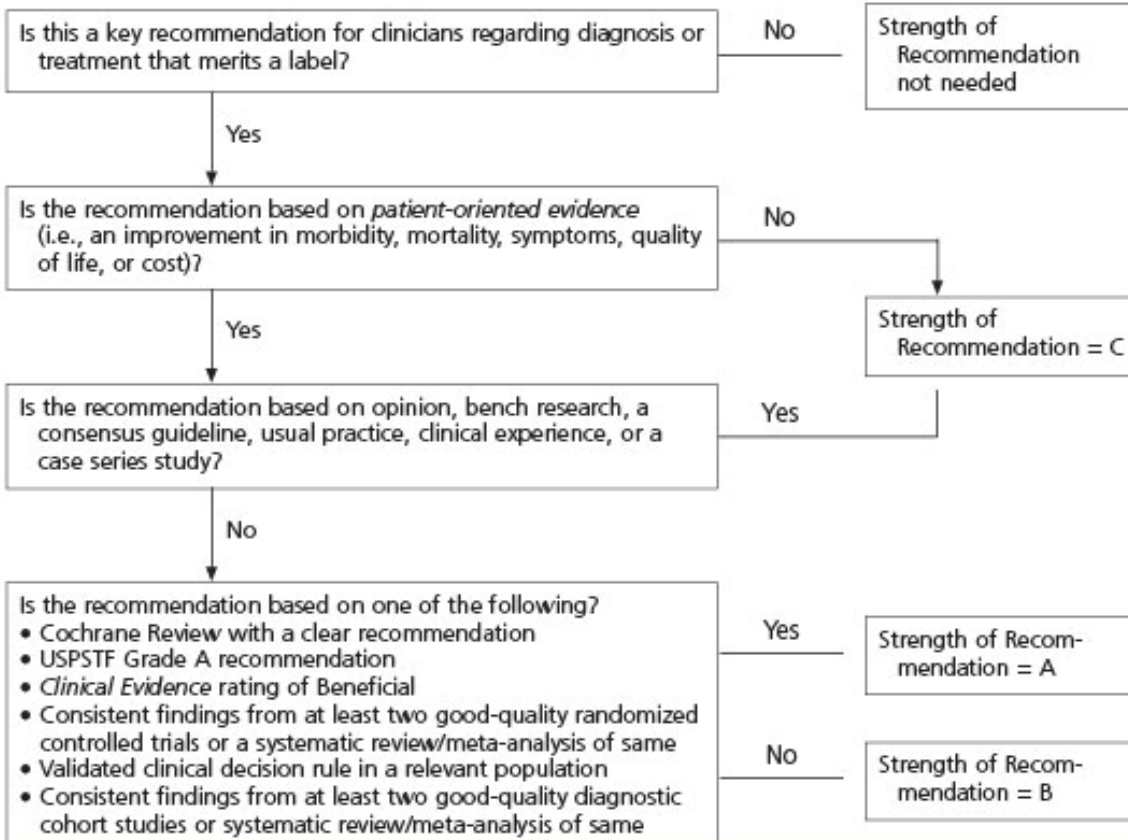
Strength of Recommendation	Definition
A	Recommendation based on consistent and good-quality patient-oriented evidence
B	Recommendation based on inconsistent or limited-quality patient-oriented evidence
C	Recommendation based on consensus, usual practice, opinion, disease-oriented evidence, or case-series for studies of diagnosis, treatment, prevention, or screening

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

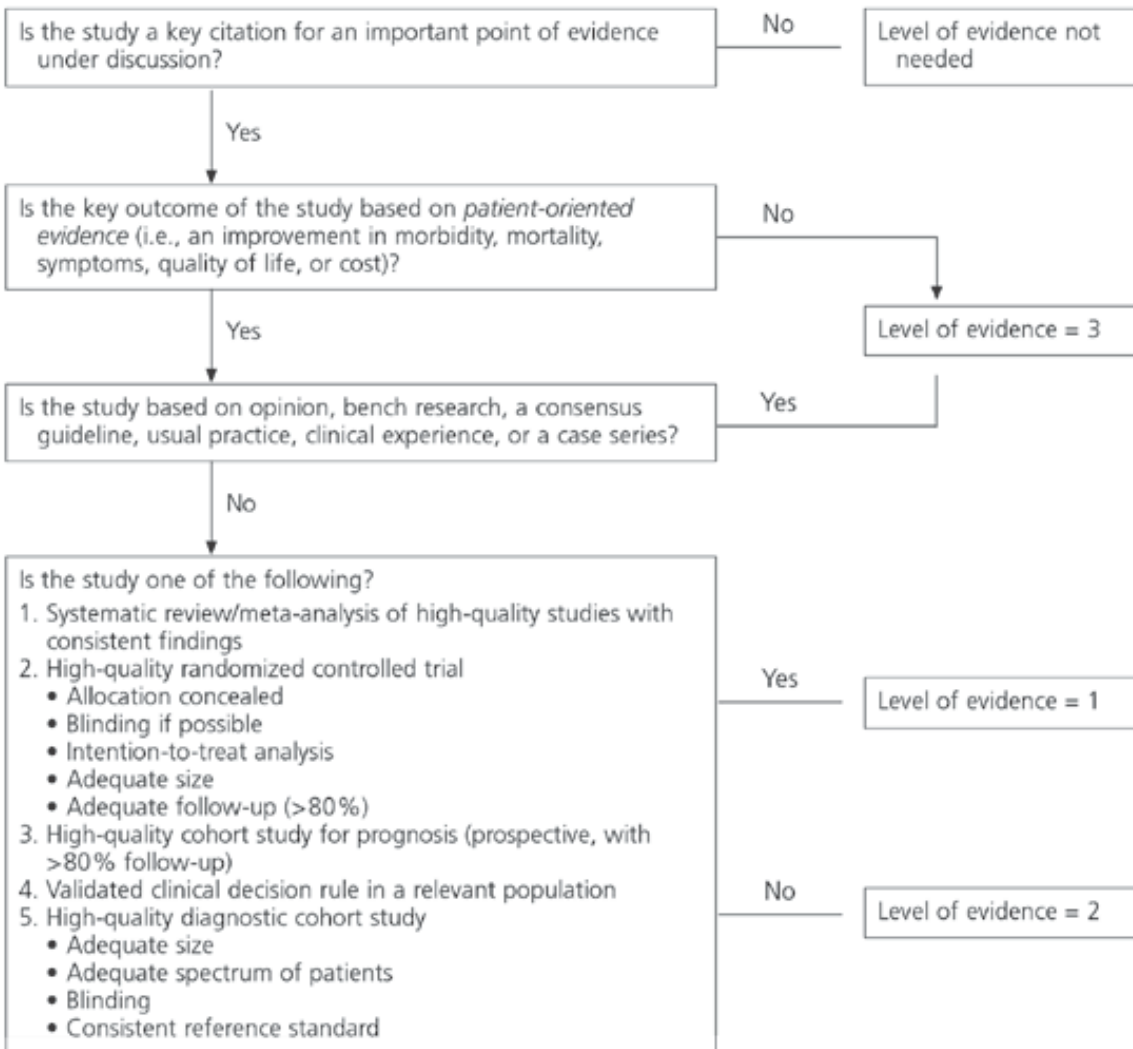
Study Quality	Evidence available for Diagnostic Tests	Evidence available for Treatment/Prevention/Screening Purposes	Evidence available for Prognostic Tests
Level 1 – Good quality patient-oriented evidence	1) Validated clinical decision rule 2) Systematic Review or Meta-Analysis of high-quality studies 3) High-quality diagnostic cohort study	1) Systemic Review or Meta-Analysis of Randomized Control Trails (RCT) with Consistent Findings 2) High-quality individual RCT 3) All-or-none study	1) Systematic Review or Meta-Analysis of good-quality cohort studies 2) Prospective cohort study with good follow-up
Level 2 - Limited-quality patient-oriented evidence	1) Unvalidated clinical decision rule 2) Systemic Review/Meta-Analysis of lower-quality studies or studies with inconsistent findings 3) Lower-quality diagnostic cohort study or diagnostic case-control study	1) Systemic Review or Meta-Analysis of Randomized Control Trails (RCT) with inconsistent findings 2) Lower-quality clinical trail 3) Cohort Study 4) Case Control Study	1) Systematic Review or Meta-Analysis of lower-quality cohort studies or with inconsistent results 2) Retrospective cohort study or prospective cohort study with poor follow-up 3) Case control study 4) Case Series
Level 3 – Other Evidence	Consensus guidelines, extrapolations from bench research, usual practice, opinion, disease-oriented evidence (intermediate or physiologic outcomes only), or case series for studies of diagnosis, treatment, prevention, or screening		

With these definitions in mind, the following algorithms can be used to determine the Level of Evidence for a body of information or for a single study

Strength of Recommendation Based on a Body of Information



Strength of Recommendation Based on a Single Study



Submitted 10-19-2020

V.4 Search Strategies-From PICO to Primary Literature (Amogh Karnik)

Once you've identified your PICO question, you may be wondering how to find research articles that relate to it.

Method 1 – The PubMed PICO Tool

A quick and easy way to find what you're looking for is by using a nifty PICO search tool at this link:

<https://pubmedhh.nlm.nih.gov/nlmd/pico/piconew.php>

Let's say you're interested in investigating the effect of using a wearable cardioverter-defibrillator (i.e., a LifeVest) compared to medical therapy for patients with ischemic cardiomyopathy after an MI. Our PICO question could be formulated as below. You can also search by specific publication type.

Search MEDLINE/PubMed via PICO with Spelling Checker

Patient, Intervention, Comparison, Outcome

go.usa.gov/xFn

Patient/Problem:

Medical condition:

Intervention:
(therapy, diagnostic test, etc.)

Compare to:
(same as above, optional):

Outcome:
(optional)

Select Publication type:

Not specified ▼
Not specified
Clinical Trial
Meta-Analysis
Randomized Controlled Trial
Systematic Reviews
Reviews
Practice Guideline

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

This is what our search results look like:

PubMed for Handhelds

US National Library of Medicine

Term: P(myocardial infarction) I(wearable cardioverter-defibrillator) C(medical therapy) O(death) RCT

2 results:

1. Wearable Cardioverter-Defibrillator after Myocardial Infarction.
Olgin JE; Pletcher MJ; Vittinghoff E; Wranicz J; Malik R; Morin DP; Zweibel S; Buxton AE; Elayi CS; Chung EH; Rashba E; Borggreffe M; Hue TF; Maguire C; Lin F; Simon JA; Hulley S; Lee BK;
N Engl J Med; 2018 09; 379(13):1205-1215. PubMed ID: 30280654
[\[TBL\]](#) [\[Abstract\]](#) [\[Full Text\]](#) [\[Related\]](#)
2. Intramyocardial transplantation of autologous CD34+ stem cells for intractable angina: a phase I/IIa double-blind, randomized controlled trial.
Losordo DW; Schatz RA; White CJ; Udelson JE; Veereshwarayya V; Durgin M; Poh KK; Weinstein R; Kearney M; Chaudhry M; Burg A; Eaton L; Heyd L; Thorne T; Shturman L; Hoffmeister P; Story K; Zak V; Dowling D; Traverse JH; Olson RE; Flanagan J; Sodano D; Murayama T; Kawamoto A; Kusano KF; Wollins J; Welt F; Shah P; Soukas P; Asahara T; Henry TD
Circulation; 2007 Jun; 115(25):3165-72. PubMed ID: 17562958
[\[TBL\]](#) [\[Abstract\]](#) [\[Full Text\]](#) [\[Related\]](#)

As you can see, the first search result is an RCT published in the New England Journal of Medicine, investigating whether the use of wearable cardioverter-defibrillators has an effect on arrhythmia-related death in patients with ischemic cardiomyopathy following an acute MI.

However, this search tool isn't perfect and may not always yield helpful results. In that case, it's best to head to PubMed.

Method 2 – Clinical Queries

PubMed has a new way of looking for clinically relevant studies that you may find useful and perhaps a bit easier to use than MeSH terms. These are called Clinical Queries.

On the PubMed home page, look for “Clinical Queries” under the PubMed Tools menu (shown below).

PubMed

PubMed comprises more than 29 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Using PubMed	PubMed Tools	More Resources
PubMed Quick Start Guide	PubMed Mobile	MeSH Database
Full Text Articles	Single Citation Matcher	Journals in NCBI Databases
PubMed FAQs	Batch Citation Matcher	Clinical Trials
PubMed Tutorials	Clinical Queries	E-Utilities (API)
New and Noteworthy	Topic-Specific Queries	LinkOut

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Enter your desired search terms. PubMed will search three categories of articles based on your search terms – Clinical Studies, Systematic Reviews, and Medical Genetics (which may not be as relevant for our purposes).

PubMed Clinical Queries

Results of searches on this page are limited to specific clinical research areas. For comprehensive searches, use [PubMed](#) directly.

myocardial infarction defibrillator

Clinical Study Categories

This column displays citations filtered to a specific clinical study category and scope. These search filters were developed by [Haynes RB et al.](#) See more [filter information](#).

Systematic Reviews

This column displays citations for systematic reviews, meta-analyses, reviews of clinical trials, evidence-based medicine, consensus development conferences, and guidelines. See [filter information](#) or additional [related sources](#).

Medical Genetics

This column displays citations pertaining to topics in medical genetics. See more [filter information](#).

After a search has been performed, there's also an option to filter clinical study results based on various categories, including etiology, diagnosis, therapy, prognosis, and clinical prediction guides. PubMed will automatically select which category is most appropriate based on your search terms, but you can select different filters if you're looking to do some more background reading about a particular topic.

Results of searches on this page are limited to specific clinical research areas. For comprehensive searches, use [PubMed](#) directly.

myocardial infarction defibrillator

Clinical Study Categories

Category:
Scope:
Results: 5
Clinical prediction guides

Wearable Cardioverter-Defibrillator after Myocardial Infarction.

Olgin JE, Pletcher MJ, Vittinghoff E, Wranicz J, Malik R, Morin DP, Zweibel S, Buxton AE, Elayi CS, Chung EH, et al. N Engl J Med. 2018 Sep 27; 379(13):1205-1215.

Impact of mineralocorticoid receptor antagonists on the risk of sudden cardiac death in patients with heart failure and left-ventricular systolic dysfunction: an individual patient-level meta-analysis of three randomized-controlled trials. Rossello X, Ariti C, Pocock SJ, Ferreira JP, Girend N, McMurray JJV, Van Veldhuisen DJ, Pitt B, Zannad F. Clin Res Cardiol. 2018 Sep 27; . Epub 2018 Sep 27.

Systematic Reviews

Results: 5 of 19

Ventricular tachycardia-inducibility predicts arrhythmic events in post-myocardial infarction patients with low ejection fraction. A systematic review and meta-analysis.

Disertori M, Masè M, Rigoni M, Nollo G, Ravelli F. Int J Cardiol Heart Vasc. 2018 Sep; 20:7-13. Epub 2018 Jun 14.

Outcomes in syncope research: a systematic review and critical appraisal.

Solbiati M, Bozzano V, Barbic F, Casazza G, Dipaola F, Quinn JV, Reed MJ, Sheldon RS, Shen WK, Sun BC, et al. Intern Emerg Med. 2018 Jun; 13(4):593-601. Epub 2018 Jan 18.

Medical Genetics

Topic:

Results: 5 of 19

Telomere shortening and telomerase activity in ischaemic cardiomyopathy patients - Potential markers of ventricular arrhythmia.

Sawhney V, Campbell NG, Brouillette SW, Coppen SR, Harbo M, Baker V, Ikebe C, Shintani Y, Hunter RJ, Dhinoja M, et al. Int J Cardiol. 2016 Mar 15; 207:157-63. Epub 2016 Jan 7.

The Relation between eNOS -786 C/T, 4 a/b, MMP-13 rs640198 G/T, Eotaxin 426 C/T, -384 A/G, and 67 G/A Polymorphisms and Long-Term Outcome in Patients with Coronary Artery Disease.

Kincl V, Máchal J, Drozdová A, Panovský R, Vašků A. Dis Markers. 2015; 2015:232048. Epub 2015 Sep 30.

Method 3 – Using MeSH Terms

If the first two methods fail, you always have the option to search through PubMed directly.

First, start by searching for the problem (P) and intervention (I) components of your PICO question. By using the “AND” operator, we can ensure that the articles contain both phrases that we're interested in.

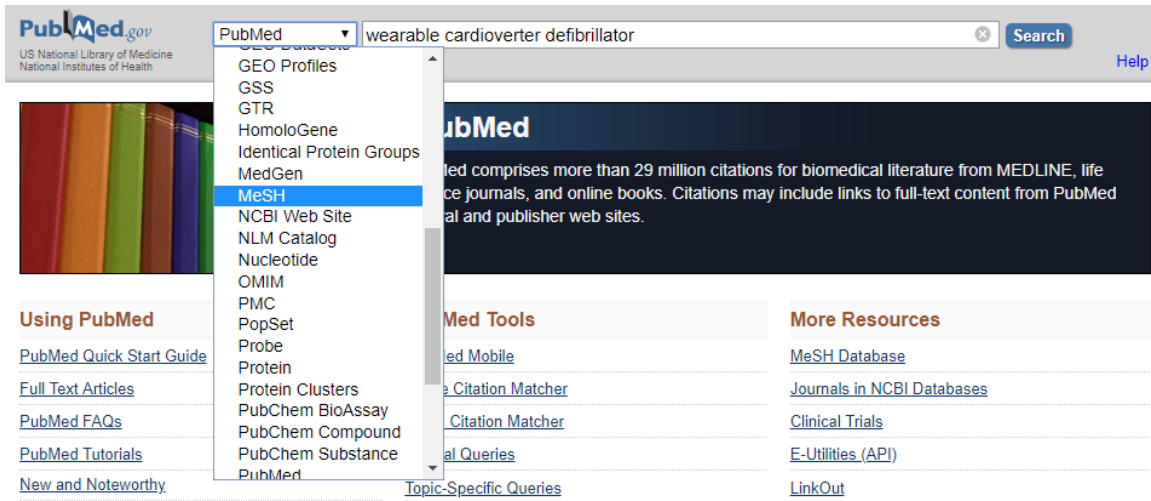
Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Next, take a look at the “Search Details” box to verify that your search query is using the appropriate MeSH terms. Remember, MeSH terms are standardized terms used in PubMed that help identify specific topics regardless of the way that they’re worded or described by authors in various articles.

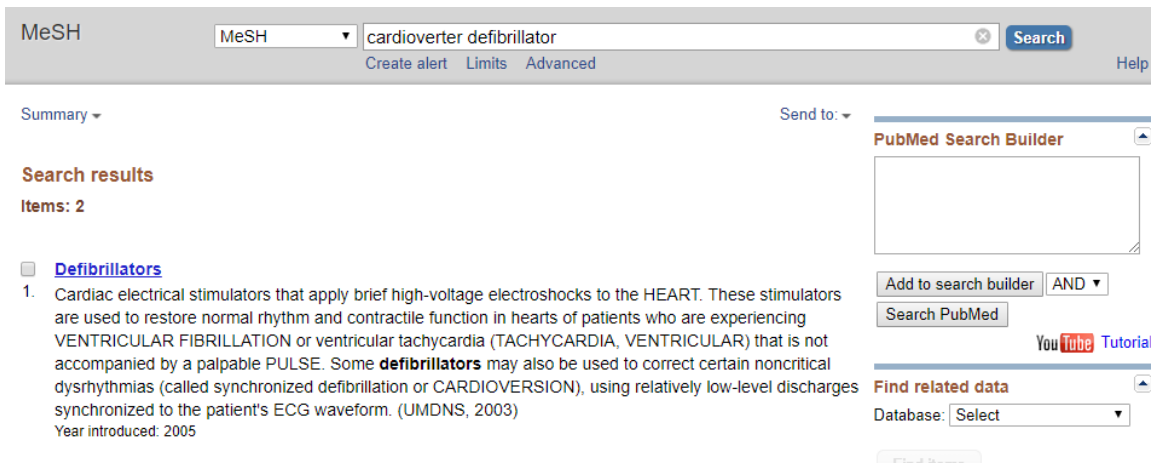
In this case, we can see that “myocardial infarction” is a MeSH term, but we weren’t able to find one that corresponds effectively with the wearable cardioverter-defibrillator. In the case where PubMed can’t automatically figure out what MeSH term to search for, we can actually search the MeSH database to identify which one will be the most appropriate.

Go back to the home page and select “MeSH” where you had originally selected PubMed as the database that you’re interested in searching.

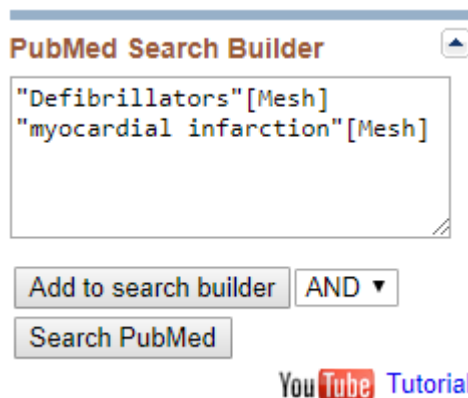
Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine



Sometimes, there aren't any MeSH terms that are compatible with our search terms (like in this case). If we drop the "wearable" portion, however, we can see a list of possible MeSH terms.



Now, we can use this menu to create a custom search using our identified MeSH terms. Click on the check box next to "Defibrillators" and click "Add to search builder". Since we already know that "Myocardial Infarction" is also a MeSH term, we can also type this into the search builder as shown.



These are our search results.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

The screenshot shows a PubMed search interface. At the top, the PubMed logo is on the left, and the search bar contains the query "Defibrillators"[Mesh] "myocardial infarction"[Mesh]. Below the search bar are links for "Create RSS", "Create alert", and "Advanced". On the right, there is a "Search" button and a "Help" link. The main content area is divided into several sections: "Article types" with a list of options like "Clinical Trial" and "Randomized Controlled Trial"; "Format: Summary", "Sort by: Most Recent", and "Per page: 20"; "Filters: Manage Filters"; "Send to" and "Sort by" buttons; "Search results" showing "Items: 1 to 20 of 716" and navigation controls; a list of search results starting with "Wearable Cardioverter-Defibrillator after Myocardial Infarction" by Olgin JE et al. (2018); and a "Results by year" bar chart with a "Download CSV" link.

If we want to narrow these results based on study type, we can click on “Customize” at the top left corner. This opens a list of checkboxes, where we can choose to filter by RCT, systematic review, meta-analysis, case-control study, and on and on.

January, 2019

V.5 Troubleshooting your search for evidence (Gwen Caffrey)

By this point, you may have constructed a PICO question to be answered, or perhaps are struggling to formulate your question without an idea of what literature already exists on a topic. You go to PubMed, type in a few key words, and are inevitably confronted with tens of thousands of search results. Or worse, you end up with only a handful of articles, none of which seems helpful. Do you take this as a sign that your question is unanswerable? Or is it possible that with a few simple steps, you can narrow your search from overwhelmingly broad to a more manageable quantity of articles? Here are a few tips for troubleshooting your initial search for evidence.

Since everyone's favorite organ is the bone marrow (right?), let's consider a search regarding red cell transfusion thresholds in adult ICU patients. If we type "transfusion" into the PubMed search bar, we come up with over 150,000 articles! You could likely spend the remainder of your residency sifting through those results and STILL fail to answer your clinical question. Let's see how we can narrow this down.

1. Consider using a MeSH term to make sure you are using terminology that PubMed recognizes as a known medical category. On the PubMed home page, click "MeSH Database" under the heading "More Resources" and type in your topic of interest. Type in "Transfusion" and hit "Search." You'll notice that this allows you to then select the type of transfusion more specifically. In this example, select "erythrocyte transfusion," then "Add to Search Builder" and finally "Search PubMed." Just like that, you've reduced the number of search results by about 95%!
2. Since most of our highest quality evidence comes from randomized controlled trials, we can narrow our search further by selecting the article type we are trying to see. On the left-hand side of the screen, under "Article Types" click "Customize," then select "Randomized Controlled Trial" from the options listed. Once it appears under the options, click it one more time to narrow your search. You'll notice we've once again narrowed our search by another 94%!
3. If you're interested in more recent articles on red cell transfusions, you can select a range of publication dates to make sure you aren't reading older evidence. Click "5 years" to narrow down our results by another 72%. You should now be looking at about 130 results.
4. Select the age range most appropriate for your search. In our case, we do this by first clicking "Show additional filters" on the left-hand side of the screen, selecting "Ages" from the options, and then "Adult: 19+ years." Depending on your clinical question, you may be interested in only geriatric patients or perhaps only young adults. PubMed includes pediatric studies, which of course are not as directly applicable to most of our clinical questions. This should bring your results down to fewer than 100 articles. We're almost done!

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

5. At this point, we can enlist the help of our **Boolean operators** to narrow things down even further. As a reminder, Boolean operators are words that we can use to either broaden or narrow down search results in electronic databases. They include AND, OR and NOT. In our particular search, go ahead and add, “AND critical” to the search bar at the top of the screen. This will largely narrow your results to those regarding critically ill patients. Take a look – we’re down to only 20 articles to look through!

No recommendations regarding troubleshooting your search would be complete without a reminder to **consult your friendly neighborhood academic librarian** for further assistance. They are generally available on a walk-in basis, but are also accessible by phone, email or by appointment. It is safe to say they are a vastly under-utilized resource when it comes to clinical queries!

Jan. 2018

V.6 An Algorithm to Assess Study Quality (Bianca Di Cocco, GSM4)

One of the reasons I decided to sign up for this elective was to get better (and faster!) at assessing whether a piece of primary literature was of good quality. My goal for this chapter is to provide a quick checklist of questions you should be asking yourself when reading through a paper reporting the results of a randomized control trial, along with the reasoning for why these questions are important. Hopefully this will help you quickly identify strengths and weaknesses of a study, and therefore help you decide whether the evidence is compelling or simply *meh*.

1. Is the study “double-blinded”?

A double-blinded study is one where both the patient and the investigator do not know who is getting the intervention as opposed to the placebo. When studies are not blinded, participants may report sensations or symptoms based on what they might know about the study medication or placebo (ex: a patient who knows they are receiving colchicine may believe they are having more stomachaches than usual). Similarly, investigators may project a benefit if they know certain patients are receiving the intervention. Subjective measures in the studies (such as surveys or questions regarding quality of life) are particularly subject to being skewed in un-blinded studies, so this is important to keep an eye on. There are some cases where blinding is impractical or immoral (such as in surgical studies); however, those studies should make an effort to obtain a third-party assessor who was not involved in the surgery or choose objective outcomes on which to base their study.

2. Are patient baseline characteristics similar between participants in all arms of the trial?

In an RCT, the goal of randomization is to reduce selection bias, and therefore ensure that both arms of the study are not skewed based on patients' baseline characteristics. Randomization should be done in a way such that investigators at a site do not know the “pattern;” for example, if a site knows that every other patient is going to receive placebo, they might “save” certain patients to enroll at a certain time, when they think they might get a treatment. If a study is using a computerized randomization system, however, we can usually assume that the randomization scheme was concealed from investigators. That said, randomization isn't always perfect, so it's important to double-check the table in each study (typically Table 1!) that lists the baseline characteristics of patients who have been randomized to each arm. You want to make sure that the percentage of patients with diabetes, for example, is similar in both groups. Otherwise, a difference between treatment arms may not actually be due to the intervention you are testing, and instead may be due to a baseline characteristic!

3. What is the percentage of women in the study? What about African Americans? What's the average age of participants?

While it's important to have similar baseline characteristics between arms of the study, it's just as important to note the actual percentages of those characteristics in the people enrolled in the study. The goal of a clinical trial is to provide us with data that we can then use to generalize to a larger population of people. Therefore, we want the study to match the *population of interest* as closely as possible. An important thing to note is that this doesn't necessarily mean the percentages should match the *general population* of the United States (ex: 50% men, 50% women). Certain diseases are skewed based on gender, age, and ethnicity, so it's important to keep that in mind (example: a study of systemic lupus erythematosus should enroll more women than men, as 90% of patients with lupus are women). Ensuring that a study's sample population somewhat matches the larger population of interest is important in ensuring generalizability of a study.

4. Where was the study conducted?

Similarly, to number 3, knowing where a study was conducted can help us determine how generalizable a study is to our desired population. If a study is conducted in China with a primarily Chinese population who have different diets, habits, and customs may make the study a little less generalizable to our patient population here in the United States. However, if the study was a large, international study, it gives it a bit more credibility as an intervention was likely tested on a large group of people with diverse backgrounds. Finally, keep an eye out for those VA studies: while we can gather a lot of great information from the VA, VA patients are a very specialized group (typically older, white men who have served in the military).

5. Is the study adequately powered/is there an adequate sample size?

Sample size calculations are discussed in more depth in another chapter ([by Haley Moulton](#)), but essentially, prior to initiating a clinical trial, investigators will calculate how many patients they need to enroll in order for the study to be adequately powered. I always like to double-check that these numbers were met, to insure that the study was not underpowered, and therefore unable to truly answer the study question.

6. Do patients stick with the trial all the way through?

Patients who do not make it to the end of the study are often called "lost to follow-up," and it's important to see how many people this has happened to. Sometimes, the outcome measure is dependent on patients making it to a certain study visit (such as Week 24), and if a lot of people drop out before reaching that date, it could affect the validity of that outcome. Also, if a lot of patients drop out, it could be a sign that the regimen is too difficult to follow, and therefore would be unlikely to be followed by patients outside of the study. Finally, if more people drop out of one arm of the study vs. the other, it's important to investigate why. Is it because of a perceived lack of benefit? Or perhaps it's due to some awful side effect that people just can't handle? All of this is important to assess.

7. **Finally: Did I read the discussion section closely?**

Almost every author should address their perception of their own study's weaknesses in the discussion section. Therefore, it's crucial to read that part of the paper closely when assessing study quality. Authors are typically experts in their field, so they may point out inadequacies that you didn't even think about! So make sure to closely read the discussion section to see what the authors think could have been done better—and then see if you agree with their rebuttals and explanations. On the other hand, looking at the primary outcomes of interest yourself should be preferable to simply accepting the authors' conclusions- that's why it's essential to drill down to the numbers. Also, if a study is a bit controversial, keep an eye out for "letters to the editor" written to the journal, where other specialists in a field might have written in with either their support or their concerns regarding the paper.

While this isn't an exhaustive list, hopefully it helps you get into the groove of assessing clinical trials and makes your process a little more stream-lined! Happy reading!

References:

1. Evans, A., & Mints, G. (2018). Evidence-based medicine. In C. Armsby (Ed.), *UpToDate*. Retrieved December 2, 2019, from <https://www.uptodate.com/contents/evidence-based-medicine>
2. Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2019). *Evidence-based medicine: how to practice and teach EBM* (5th ed.). Edinburgh: Elsevier.

12/4/2019

V.7 Critical Appraisal for Randomized Controlled Trial (Joan Chandra)

Throughout our careers as clinicians, we will be presented with studies that will shape the way we practice medicine. Critical appraisal is necessary to glean the information that we need to decide whether or not we incorporate new evidence into our practice. These questions are meant as a guide through the assessment of an RCT.

SCREENING

- Does the study question match your question?
- Was the study design appropriate?

VALIDITY

Patient Follow-Up

- ☐ Were all patients who entered the trial properly accounted for at its conclusion?
Losses to follow-up should be less than 20% and reasons for drop-out should be given.
- ☐ Was follow-up long enough?

Randomization

- Were the recruited patients representative of the target population?
- Was the assignment of patients to treatment randomized and concealed?

Intention to Treat Analysis

- Were patients analyzed in the groups to which they were randomized
- Were all randomized patient data analyzed? If not, was a sensitivity or “worst case scenario” analysis done?

Similar Baseline Characteristics of Patients

Were groups similar at the start of the trial?

Blinding

- Were patients, health workers, and study personnel “blind” to treatment?
- If blinding was impossible, were blinded raters and/or objective outcome measures used?

Equal Treatment

- Aside from the experimental intervention, were the groups treated equally?

Conflict of Interest

- ☐ Are the sources of support and other potential conflicts of interest acknowledged and addressed?

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Summary of Article's Validity

- Notable study strengths, weaknesses, or concerns?
- How serious are the threats to validity and in what direction could they bias study outcomes?

CLINICAL IMPORTANCE

- How large was the treatment effect? Use EER, CER, RRR/RRI, ARR/ARI, and NNT/NNH to help with your evaluation.
- How precise was the treatment effect? Evaluate the confidence interval as well as the p-value which can relay the statistical significance.

References:

1. "Expanded Critical Appraisal Worksheet with Key Learning Points". Duke Program on Teaching Evidence Based Practice.
2. "Critical Appraisal Form for Single Therapy Studies". Oxford Centre for Evidence-Based Medicine.

9-18-2018

V.8 Comparing and Contrasting Two or More Studies (Susan Wang)

While analyzing papers, you will inevitably come across multiple studies that may have different conclusions. While meta-analysis may be useful in using statistical analysis to combine existing randomized control trials, the result of this analysis is only as good as the trials which you used for this analysis. Consider the following approach when attempting to analyze multiple bodies of evidence. In addition, refer to chapter 10, on further critical appraisal for randomized controlled trial.

Design:

- What type of study is this? (RCTs, observational, etc.)
- Was this a single center or multicenter study?
- Was this study blinded or un-blinded?

Population:

- How many people were in this study?
- What was the average age, age range, gender ratio, etc.?
- What types of patients were in the study? Surgical? Medical?
- Where was this study performed?
- What were the inclusion/exclusion criteria in this study?

Intervention versus Control:

- What is the author's definition of the intervention versus the control? Be as specific as possible.
- How many patients actually received the intervention versus the control? Is this an as treated or intention to treat analysis?

Results:

- What were the results of this study?
- What are the important statistical analyses for this result?

For example, here is a comparison of two different randomized control trials published in JAMA and NEJM respectively. The question being asked was, "In critically ill patients with significant renal impairment, is there a mortality benefit to instituting early renal replacement therapy compared to a strategy of delayed RRT?" A side-by-side comparison allows the reader to judge whether these two studies were comparable, whether it would be reasonable to include them in a meta-analysis, and what were some of the differences between them

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

		ELAIN (JAMA)	AKIKI (NEJM)
Design		Randomized, single-center, <u>unblinded</u> , controlled trial	Randomized, multicenter, <u>unblinded</u> , controlled trial
Population		231 predominantly surgical patients in ICU in Germany	620 predominantly medical patients in 31 ICUs in France
	Inclusion criteria	KDIGO Stage 2: Cr >2 times baseline or UOP <0.5ml/kg/hr for >12 hours 1 of following: severe sepsis, use of <u>norepi</u> /epi, refractory fluid overload, <u>nonrenal organ dysfunction</u>	KDIGO Stage 3: Cr >3 times baseline, UOP <0.3ml/kg/hr for >24hrs or anuria for >12 hours 1 of following: mechanical ventilation or use of <u>norepi</u> /epi
	Exclusion criteria	Baseline GFR <30, <u>prev</u> RRT, AKI caused by occlusion or lesion of renal artery, glomerulonephritis, interstitial nephritis, vasculitis, <u>postrenal</u> obstruction, TTP-HUS	BUN >112, K >6, pH < 7.15, pulmonary edema with O ₂ >5L/min or <u>mech</u> vent with FIO ₂ > 50%
Baseline SOFA (early vs delayed)		15.6 vs 16.0	10.9 vs 10.8
Intervention	Early RRT	Within 8 hours of KDIGO 2	Within 6 hours of KDIGO 3
	Method of RRT	CVVHDF	Physician-Dependent (55% intermittent HD)
Control	Delayed RRT	Within 12 hours of KDIGO 3	If develop Urea > 40mmol/L, K>6, pH <7.15, acute pulmonary edema, oliguria/anuria at >72 hours
	% of patients who received RRT	91% at mean 25.5 hours	51% at mean 57 hours
Results	Mortality (early vs delayed)	39.3% vs 54.7% at 90 days (HR: 0.66, CI: 0.45 to 0.97)	48.5% (CI 42.6-53.8) vs 49.7% (CI 43.8-55.0) at 60 days
	RRT (early vs delayed)	At 90d, 13% vs 15% (OR: 0.87, CI: 0.31 to 2.44)	At 60d, 2% vs 5%

Note that there are many differences in the definition of the method as well as the inclusion and exclusion criteria. This is likely responsible for the different results obtained. From this example, you can observe that these two studies would not lend themselves well to a meta-analysis because their methodology differed greatly. Therefore, a careful comparison and contrast between the articles is necessary before attempting any meta-analysis.

V.9 Assessing the Risk of Bias of Randomized Controlled Trials in Systematic Reviews and Meta-Analyses (Chris Lindholm)

The Cochrane risk of bias tool is a tool that was constructed to evaluate and assess for the risk of bias in studies included in a systematic review. The tool identifies studies that are at high risk of having bias, low risk of having bias or unclear risk of bias. To achieve this determination, 7 domains are analyzed – sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective outcome reports and other issues.

Sequence generation refers to the method used to produce comparable groups, the sufficiency to which the method will produce comparable groups and the author's description of the method to allow an assessment of whether or not the method should produce comparable groups. This is to help determine **selection bias**.

Allocation concealment refers to the method used to conceal the allocation sequence, whether or not differences in allocation could have been foreseen before or during enrollment and the author's description of the methods used to conceal allocation. This also helps to determine **selection bias**.

Blinding of participants and personnel refers to all measures to blind the study participants and study personnel from knowledge of which intervention the study participant received and the author's description of the blinding to allow for an assessment. This is to help determine **performance bias**.

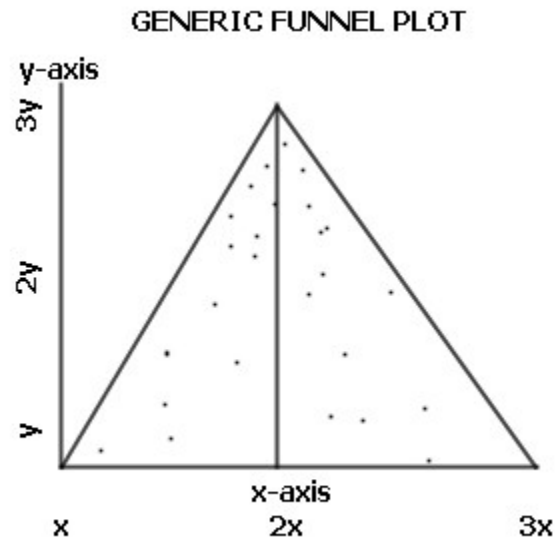
Blinding of outcome assessment refers to all measures used to blind the outcome assessors from knowledge of which intervention the study participants received and the author's description of the blinding to allow for this determination. This is to help determine **detection bias**.

Incomplete outcome data refers to the completeness of the outcome data for each described outcome, describing how patients were excluded and reported and the number of patients in each intervention group and those that were excluded. This helps to determine **attrition bias**.

Selective outcome reports refer to the assessment of outcomes by the authors, the possibility of selectively reporting outcomes and the author's description of this to allow for assessment. This helps to determine **reporting bias**.

Other issues refer to any other important concerns leading to bias that are not covered by the above measures which helps to determine other types of bias.

Funnel plots are graphs that are constructed to assess for **publication bias**. They are scatter plots of treatment effects against precision (see figure). As a result of this, studies with high precision will fall near the median, and if there is no publication bias, theoretically studies with lesser degrees of precision will fall equally on both sides. If there is no publication bias, the scatter-plot should look somewhat symmetrical. Asymmetric scatter plots suggest publication bias exists and further work can be done looking into why that may have occurred (ex: lack of small negative trials).



References:

1. <https://methods.cochrane.org/bias/resources/cochrane-risk-bias-tool>
2. https://en.wikipedia.org/wiki/Funnel_plot

November 2018

V.10 Systematic Reviews and Assessing the Quality of Evidence with GRADE (Briana Goddard, GSM4)

What is a Cochrane Review?

As described in the chapter “[Systematic Reviews and Meta Analyses](#)” by Alex Donovan, a systematic review answers a question by summarizing the evidence available meeting certain inclusion requirements. A Cochrane systematic review is one that meets strict criteria for research and reporting methods. Cochrane is a global network made up of health care practitioners, researchers, and patient advocates. The mission of the organization is to produce high quality systematic reviews in order to promote evidence-based decisions. In order to publish a Cochrane Review, the study must be registered with Cochrane and have a written protocol before it is begun. If accepted, Cochrane will then provide a team to help ensure that the authors adhere to the high standards of methodology.

When discussing systematic reviews, we mainly focus on systematic reviews of interventions. However, Cochrane publishes five different types of reviews, with a separate approach to each: reviews of the effects of interventions, reviews of diagnostic test accuracy, reviews of prognosis, overviews of reviews, and reviews of methodology.

What Type of Question does a Systematic Review Answer?

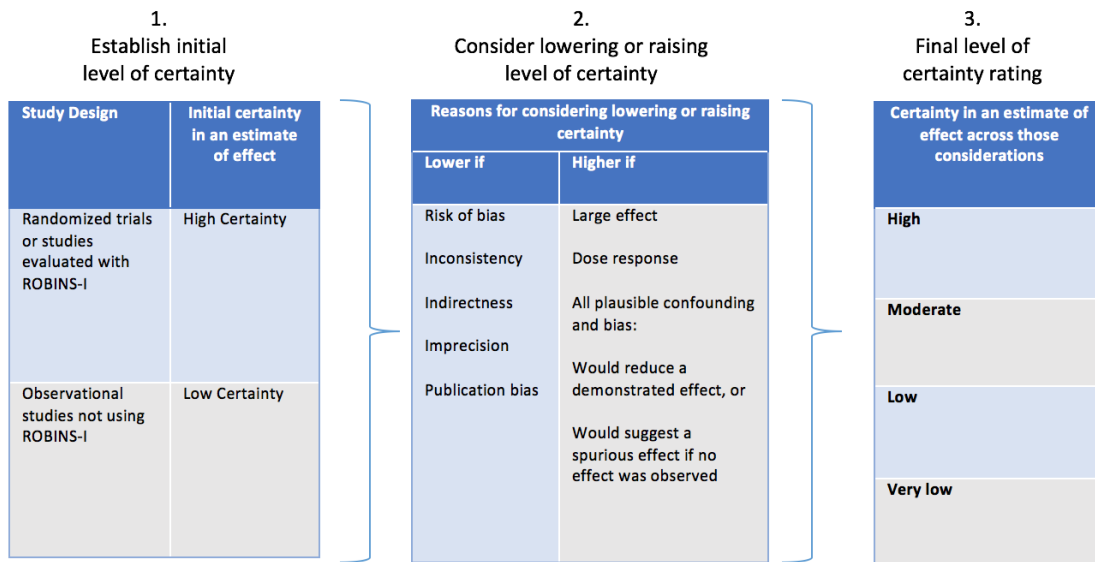
The goal of a systematic review is to provide a health care practitioner with evidence to answer a clinical question. The question answered in a systematic review should meet the “FINER” criteria. **F – Feasible.** A question is feasible if it can be answered using the evidence available. The question should be narrow enough as to not include an unmanageable amount of data, but broad enough that a number of trials meet eligibility requirements. **I – Interesting.** Systematic reviews require a significant amount of time, so authors should be interested in the topic they will be researching. **N – Novel.** Reviews should address gaps in knowledge. It is helpful to be aware of pre-existing reviews that have attempted to answer the same or similar questions. Ongoing reviews can be searched in the PROSPERO registry. **E – Ethical.** Ethical considerations of a review may include costs, the framing of results, and the potential implications of results. **R – Relevant.** Reviews are relevant if health care providers can use the information provided to make clinical decisions. The reporting of reviews should be transparent so that readers can assess the evidence and determine how it will affect their clinical decision making.

What Studies Are Included in a Systematic Review?

Each systematic review will have eligibility criteria that studies must meet to be included. These eligibility criteria are based upon the PICO question that the review is attempting to answer. The population, intervention, and comparison components of the question form the basis for the eligibility criteria. Authors also must consider what types of studies to include in their review. Randomized trials should be included if they are feasible for the intervention of interest. Non-randomized trials may be included if randomized trials are unable to address the intervention of interest, or for interventions that cannot be randomized. Finally, studies, not reports of studies, are the unit of interest in systematic reviews. Therefore, if multiple reports of the same study are found, then the reports should be collated.

How Do You Assess the Quality of Evidence in a Systematic Review?

The quality of evidence included in a systematic review greatly affects the quality of the review itself. The certainty of the evidence from the included studies is assessed using the GRADE approach. The Grade of Recommendation, Assessment, Development and Evaluation Working Group (GRADE Working Group) developed a system for grading the certainty of evidence. In the GRADE system, evidence from randomized trials begins with a high-certainty rating and is downgraded for any concerns. A non-randomized trial begins with a low certainty rating due to the potential confounding and selection bias created by the lack of randomization. The GRADE system is summarized in the figure below.



The domains that may result in a downgrade of certainty are risk of bias, inconsistency, indirectness, imprecision, and publication bias. For more information about assessing the **risk of bias** and **publication bias**, refer to the chapter [“Assessing risk of bias of randomized controlled trials in systematic reviews and meta-analyses”](#) by Chris Lindholm. **Inconsistency** refers to unexplained heterogeneity, or unexplained differences in the results of studies. **Indirectness** results when a study addresses only a part of the main question being asked in the review, in terms of the population, intervention, or comparator. For example, if a review sought to evaluate an intervention for the prevention of heart disease, a study would be considered indirect if it only evaluated the intervention in patients with diabetes. **Imprecision** occurs when studies include a small population and is indicated by wide confidence intervals.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

The domains that may upgrade the certainty of evidence in non-randomized trials are large effects, dose response, and plausible confounding. **Large effect** refers to studies that show a consistent and precise large magnitude of effect. If a large effect is demonstrated, and repeat measures find a similarly large effect, investigators can be more confident that this is a real effect rather than a result of confounding or bias. Similarly, if a **dose response** gradient is present, it increases the confidence that the results of the study are accurate. The lack of randomization of trials results in confounding and selection bias. However, if the **confounding** factors work to under-estimate an effect then it may increase confidence in results, rather than decrease them. For example, “if only sicker patients receive an experimental intervention or exposure, yet they still fare better, it is likely that the actual intervention or exposure effect is larger than the data suggest.” By considering both reasons to downgrade and upgrade the certainty of evidence, investigators can come to a final estimate of certainty in the evidence of very low, low, moderate, or high.

The ROBINS-I is the Risk of Bias In Non-randomized Studies – of Intervention. This tool allows non-randomized studies to be critically appraised for their risk of bias. It covers seven domains: bias due to confounding, selection of participants, classification of interventions, deviations from intended interventions, missing data, measurements of outcome, and selection of the reported results. By analyzing these areas, investigators can categorize non-randomized trials as having a low risk of bias, a moderate risk of bias, a serious risk of bias, or a critical risk of bias.

Finally, accessing these tools is possible through the EBM Resources tab in the EBM Database, as well as through the Biomedical Library site at Dartmouth-Hitchcock.

References:

1. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
2. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919.

Submitted 4/24/2020

V.11 Questioning Quality of Qualitative Research (Sarah Baranes, GSM4)

Have you ever tried to follow a treatment algorithm and found yourself off the grid? Patients do not always fit into discrete descriptions or follow the linear trajectory from one box to the next. When we don't know why, qualitative research can uncover why data does not match the reality before us and can reveal how to proceed to realign our practice with evidence.

This chapter will attempt to briefly justify the utility of qualitative research in healthcare sciences and provide a process to help you evaluate whether the findings of qualitative studies merit integration into your evidence-based practice. This chapter will not outline methods of qualitative research because there are entire books written on the subject, and, frankly, you don't need to know that level of detail to benefit from what qualitative research has to offer.

I. Qualitative Research: An Integral Part of the Process

Quantitative research has a well-earned home in the physical sciences and its methodologies have proved indispensable to the field of medicine. The design of the randomized control trial was born out of a medical experiment testing the efficacy of streptomycin on pulmonary tuberculosis in 1946, and it has continued to be thought of as the gold standard for therapeutic research.¹

Origins of qualitative research on the other hand are rooted in social sciences with tools that were initially designed to observe and explain human behavior. Qualitative methods only appeared in health research in the last five-or-so decades and have provided insight into human behavior, patients' experiences with illness, dynamics of interprofessional teams and other nuanced variables that affect healthcare delivery.^{2,3}

¹ Bhatt A. Evolution of clinical research: a history before and beyond James Lind. *Perspect Clin Res.* 2010;1(1):6-10

² Al-Busaidi ZQ. Qualitative research and its uses in health care. *Sultan Qaboos Univ Med J.* 2008 Mar;8(1):11-9. PMID: 21654952.

³ Chafe R. The Value of Qualitative Description in Health Services and Policy Research. *Health Policy.* 2017;12(3):12-18.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

One simple way to conceptualize how qualitative methodology fits in with that of quantitative research is to think of qualitative research as hypothesis generating and quantitative research as hypothesis testing. This relationship is not new to the physical sciences; Newton posited his second law about the relationship between mass and acceleration in 1687 based on observations which were not experimentally demonstrated until a century later with the advent of the Atwood machine. In his analysis of the evolution of science, scholar Thomas Kuhn argues, "Large amounts of qualitative work have usually been prerequisite to fruitful quantification in the physical sciences."⁴

There is ongoing debate about the order of operations from hypothesis generation to testing, but for our purposes, we can accept that most areas of health sciences inquiry will involve a combination of methodologies that create a feedback loop of data production revealing new avenues for exploration.⁵ COVID-19 vaccine rates provide a pretty good example of how this cyclical process works.

Quantitative research led the fastest vaccine development in human history, but as of November 2021, 68.8% of eligible individuals in the U.S. were fully vaccinated, however, the range was wide from 72% in Vermont to 41% in West Virginia.⁶ Data indicates rural dwelling, Black, and Latinx individuals have lower vaccine rates than urban and White counterparts, but these statistics leave openings for biased interpretation. (Indeed, the thoughtful reader is already objecting to the social construct of race and searching for more meaningful categories to explain societal differences!) Qualitative research methods offer systematic methods for bridging the numbers to realities on the ground.⁵

Balasuriya et al. contextualized this data with multiple multilingual focus groups that sought to identify themes explaining acceptability and accessibility of the COVID-19 vaccine in Black and Latinx communities in New Haven, CT. A theme of pervasive mistreatment of Black and Latinx communities and associated distrust emerged and was used to identify paths for targeted interventions such as identifying and employing trusted members of the community to help disseminate information to people who were mistrustful of currently available sources.⁷

Generating appropriately nuanced solutions requires appreciating that people's unique life experiences are often resistant to generalization. Although time consuming, qualitative research involves systematic methodologies that preserve narratives of people to draw context-informed conclusions that can inform next steps. Without expertise, qualitative methodology—like statistical analysis—can be difficult to assess, but the following section will offer a stepwise approach to help you determine if the study at hand is implementing the tools of qualitative research responsibly.

⁴ Kuhn, TS. (1961). "The Function of Measurement in Modern Physical Science". *Isis*. **52** (2): 161–193 (162). JSTOR.

⁵ Graf M, Tuly R, May S. The complementary relationship between quantitative and qualitative research methods in enhancing understanding of treatment decisions, outcomes, and value assessment. *J Clin Pathways*. 2021;7(6):20-23.

⁶ CDC. Vaccination Delivery and Coverage by State in the U.S. Accessed November 15, 2021. https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-total-admin-rate-total.

⁷ Balasuriya L, Santilli A, Morone J, et al. COVID-19 Vaccine Acceptance and Access Among Black and Latinx Communities. *JAMA Netw Open*. 2021;4(10):e2128575.

II. Quality Control

The following stepwise approach integrates the simplified approach offered in Straus et al.'s *Evidence-Based Medicine: How to Practice and Teach It*⁸ with modifications based on alternative approaches found in the literature. When deciding to apply the findings of a qualitative research study to your evidence-based practice, try answering the following questions:

- (1) Does this study apply to my patient?
- (2) Were the methods of data collection explicit and appropriate?
- (3) Are the results valid and important?

Step 1: Does this study apply to my patient?

Before diving into the details of how the study was conducted, we can save ourselves time by asking if the study is relevant to our patients. Unlike the randomized sampling techniques used to gather representative study populations in RCTs, qualitative research begins with *purposeful* sampling, or identifying specific groups of people with characteristics or lived experiences that are relevant to the phenomenon being studied. The study may aim to maximize or minimize internal variability of these study groups. Both approaches necessitate comprehensive justification and explanation in a study's methodology to allow the reader to critically evaluate how applicable the findings of the study will be to his or her patient.²

While evaluating the characteristics of the sample population, you should ask if these people were the most appropriate group to answer the question that the study set out to answer.⁸ If, for example, a study wanted to explore how the community in the Upper Valley views the services offered at DHMC's emergency department but polled patients in the ED who were visiting from out of town, that's a red flag. If the characteristics of the study population are not provided clearly enough for you to make this call, toss the study.

The process of evaluating the relevance of a cohort is comparable for qualitative and quantitative research, so if we decide that the qualitative study is relevant, we can move onto step 2 and critically appraise the methodology. Fair warning: the methodologies of qualitative research may involve steps that are unacceptable in quantitative research and thus warrant adjustment of our expectations.

⁸ Straus S, Glasziou P, Richardson S, Haynes BR. *Evidence-based medicine: How to practice and teach it*. 5th Edition (updated May 2018). Elsevier, 1997

⁸ Hannes K. Chapter 4: Critical appraisal of qualitative research. *Supplementary Guidance for Inclusion of Qualitative Research in Cochrane Systematic Reviews of Interventions*. Version 1. Cochrane Collaboration Qualitative Methods Group, 2011.

Step 2: Were the methods of data collection explicit and appropriate?

Let's first talk about objectivity. Quantitative research employs randomization, blinding, matching, and statistical analysis to root out subjectivity and bias, whereas qualitative research is both defined and enriched by opinion and perspective. This is evidenced by the methods of data collection, which most commonly include observations, interviews, focus groups, and document reviews.² Some of these tools seek out opinion, and all are subject to interpretation in the analysis phase.

Astute critics may ask, what is the difference then between anecdotal (i.e., unreliable) evidence and qualitative research? The strength and credibility of qualitative research lies in the analysis of findings and the commitment to examining counter explanations.⁹ So, how do we make sure we aren't just blindly accepting an individual researcher's opinion? The first way is to weigh the conclusions considering the researchers' stated biases. If the researchers do not disclose these biases and their objectives clearly, toss the study.

We can also critically evaluate what tools the researchers use to evaluate the data. Some qualitative research organizes methodology around specific frameworks, however more often methods will pull elements from various approaches to tailor the study to the research question. As previously mentioned, do not toss a qualitative study on account of its flexible methodology. Do, however, toss it if the researchers do not provide adequate justification for the approach(es) used.

Evaluating qualitative methodology does not require familiarity with every framework of qualitative research, just as interpreting the conclusions of a RCT does not require a comprehensive understanding of cox proportional hazards models or propensity score matching (although that can help!). With a little critical thinking, we can determine if the justification for use of the employed methodology logically follows the definition of the methods provided.

Just as quantitative research can improve its credibility by involving statistician consultants, qualitative research teams often involve co-investigators or consultants with expertise in qualitative methodology to provide technical and theoretical appraisal of methods and paradigms.⁸ While mention of a professional stranger does not buy total credibility, it may weigh into your overall evaluation.

Step 3: Are the results valid and important?

In evidence-based practice, we seek evidence from research that is both internally and externally valid; the process of data collection and interpretation is reproducible, and the conclusions are generalizable.

⁹ Green J, Britten N. Qualitative research and evidence based medicine. *BMJ*. 1998;316(7139):1230-1232.

Quantitative research strengthens internal validity when multiple reviewers of the data independently draw the same conclusions from the gathered evidence. Multiple reviewers of qualitative data, on the other hand, often generate alternative interpretations.⁸ Deviant case analysis is one analytical tool that seeks out contradictory themes emerging from anomalous results in a data set and then refines the explanations to encompass all findings.⁹ While drawing multiple conclusions from the same data does not inspire confidence in the reproducibility of results, it does provide an additional dimension of understanding that can further inform theory development and thus strengthen the explanatory power of the research.

Lack of internal validity does not necessarily compromise the value of qualitative research, but lack of external variability is a different matter. Well-designed randomized control trials bolster the strength of their conclusions with larger sample sizes. Unfortunately, the time-consuming nature of qualitative methods of data collection such as interviews and focus groups often do not lend themselves to strength in numbers.

The strength and applicability of a qualitative study's results are contingent on their interpretation, which as discussed above, introduces subjectivity and bias. Methodologic rigor can minimize, or at least clarify, how bias affects results but may come at the cost of sample size. One way qualitative studies clarify what tradeoffs were made between depth and breadth in research design is to conduct sensitivity analyses, which, simply put, examine what happens to results when high- or low-quality studies are removed.² For a more sophisticated explanation of sensitivity analysis, refer to Dr. Freed's chapter, *Sensitivity Analysis in Clinical Trials*, in this EBM guide.

III. How to Apply High Quality Qualitative Evidence in Clinical Practice

We have entered an era of exponential growth of available information; whether attempting to understand the impact of the individual genome or the sociologic forces of structural violence, preserving nuanced truths is essential for developing solutions that accommodate the complexity of our reality.¹⁰ Qualitative research provides a process of systematically reviewing and collating diverse opinions, beliefs, and perspectives, bridging this growing body of evidence to the individual in front of us. While it may not be a straight line from either qualitative or quantitative research to healthcare decisions made by clinicians or patients, the aggregate effect of rigorous study tends to create more clarity and accountability for the decisions that we make. This is one of the cornerstones of evidence-based medicine.

References:

- ¹ Bhatt A. Evolution of clinical research: a history before and beyond James Lind. *Perspect Clin Res.* 2010;1(1):6-10

Mays N, Pope C. Qualitative research in health care. Assessing quality in qualitative research. *BMJ.* 2000;320(7226):50-52.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

² Al-Busaidi ZQ. Qualitative research and its uses in health care. *Sultan Qaboos Univ Med J*. 2008 Mar;8(1):11-9. PMID: 21654952.

³ Chafe R. The Value of Qualitative Description in Health Services and Policy Research. *Health Policy*. 2017;12(3):12-18.

⁴ Kuhn, TS. (1961). "The Function of Measurement in Modern Physical Science". *Isis*. **52** (2): 161–193 (162). JSTOR.

⁵ Graf M, Tuly R, May S. The complementary relationship between quantitative and qualitative research methods in enhancing understanding of treatment decisions, outcomes, and value assessment. *J Clin Pathways*. 2021;7(6):20-23.

⁶ CDC. Vaccination Delivery and Coverage by State in the U.S. Accessed November 15, 2021. https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-total-admin-rate-total.

⁷ Balasuriya L, Santilli A, Morone J, et al. COVID-19 Vaccine Acceptance and Access Among Black and Latinx Communities. *JAMA Netw Open*. 2021;4(10):e2128575.

⁸ Straus S, Glasziou P, Richardson S, Haynes BR. *Evidence-based medicine: How to practice and teach it*. 5th Edition (updated May 2018). Elsevier, 1997

⁸ Hannes K. Chapter 4: Critical appraisal of qualitative research. *Supplementary Guidance for Inclusion of Qualitative Research in Cochrane Systematic Reviews of Interventions*. Version 1. Cochrane Collaboration Qualitative Methods Group, 2011.

⁹ Green J, Britten N. Qualitative research and evidence based medicine. *BMJ*. 1998;316(7139):1230-1232.

¹⁰ Mays N, Pope C. Qualitative research in health care. Assessing quality in qualitative research. *BMJ*. ¹⁰ 2000;320(7226):50-52.

Submitted 11/2021

V.12 Measures of Impact (John Hon, GSM4)

How do you know if a particular journal or author you're citing is reputable? There are several metrics used to determine whether the article you are reading is from a "high-impact" journal or author. Here are a few well known examples, but do not limit your metrics to these specific methods- different methods of analysis are more appropriate for determining quality depending on the circumstances.

A. Citation Analysis – Impact Factor

Created by Eugene Garfield in 1961, this was one of the earliest methods of determining the impact of a journal. This is calculated in any given year via: # of instances where article was cited divided by the total number of articles published by the journal. Although widely used, it is not without its fair share of criticisms that it's not very applicable across disciplines where trends in publishing differ greatly, and it also is heavily subject to things such as editorial policies more so than the perceived "quality" of research.

As an example:

A = Total citations in 2002 to items published in *Journal X*
B = 2002 citations to items published in *Journal X* in 2000–01 (subset of **A**)
C = Number of substantive articles published in *Journal X* in 2000–01

$$\text{Impact factor} = B/C$$

Example:

Assume that in 2002, there were 3,200 citations to items published in *Journal X*. Of these, 550 were citations to items published in *Journal X* in 2000 and 2001. During those two years, *Journal X* published 72 articles.

$$2002 \text{ Journal X impact factor} = 550/72 = 7.64$$

Adapted from: GARFIELD E. The impact factor. [Internet]. *Curr Contents* 1994 Jun 20;(25):3–7. [cited 16 Aug 2002]. <<http://sunweb.isinet.com/isi/hot/essays/journalcitationreports/7.html>>.

Figure 1 Calculating impact factor

B. H-Index

Created in 2005 by Jorge E. Hirsch, this is a method of determining the impact and productivity of a particular author. This is based on a function that is related to both an author's citations in other journals and most cited papers. Although being seen as addressing many of the concerns with impact factor- there are still faults with the h-index in that it loses its accuracy across disciplines and that it is still subject to skewed data such as self-citations. Several variations of the h-index have been created in order to address many of these concerns.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

There are many other methods of citation analysis and it is pertinent to note that although widely used methods of deriving impact are easier to use and are more broadly applicable, one cannot rule out finding a more specific means of impact to determine the quality of citations. With the increasing accessibility to peer-reviewed research, there has also been a rise in the spread of “questionable” research studies and “fake news.”

A journal's impact within clinical medicine depends largely on its importance to practitioners, most of whom never write manuscripts for publication and thus never have a chance to “vote.” Citation frequency may therefore better reflect the importance of clinical journals to researchers than practitioners. Because the opinions of both practitioners and researchers are relevant in judging the importance of clinical journals, the validity of impact factor as a measure of journal quality in clinical medicine is uncertain.

Submitted 4/2019

V.13 History, Ethics, and Current State of Pediatric Research (Hira Haq, GSM4)

Medical research in children began as early as the 18th century. When a scientist named Jenner observed that exposure to cowpox seemed to offer immunity against smallpox, he created the first experimental vaccination and administered it to his own 1-year-old son. In the 19th century, the first anti-rabies vaccine was created by Louis Pasteur and administered to a 9-year-old boy who was bitten by a rabid dog. As pediatric research began to grow, so did the restrictions imposed on it. The first documented set of restrictions were not seen until the 1900's however. In 1931, the Reich Health Council, based in Germany, issued "Regulations on New Therapy and Human Experimentation" after the death of 75 children who were given experimental anti-tuberculosis vaccinations. These regulations set forth the need for consent.

As the decades went on, more guidelines were laid out, but these were driven primarily by research in adults. Most notably was the 1947 Nuremberg code, adopted after unethical studies were carried out in Nazi concentration camps. The Nuremberg Code was the first international code of research ethics that required informed consent to be obtained from all participants in human experimentation. The Nuremberg Code, however, did not directly address children. This changed in 1964 with the Declaration of Helsinki, laid out by the World Medical Association, which emphasized the well-being of research subjects over the interests of science and society. Regarding children, the Declaration stated that research was allowed as long as "permission from the responsible relative replaces that of the subject", implying the need for parental consent.

These guidelines were being set forth on an international level, but many US researchers felt that specific guidelines were needed for the US, particularly after a few high-profile studies were shown to endanger the lives of children. Most notable of these were the Fernald and Willowbrook school studies. In Fernald, a residential institution in Waltham Massachusetts for children described as "mentally retarded", children were fed radioactive iron and calcium in their cereal. Parental "permission" was obtained through a letter that stated, "We are considering the selection of a group of our brighter patients ... to receive a special diet rich in the above-mentioned substances for a period of time". No details about the diet were given, and the letter seemed to imply that the diet would benefit the children, when in reality it did not. In another study done in the 1950's, children at Willowbrook, an institution for mentally retarded children in Staten Island NY, were infected with strains of hepatitis in an effort to study the natural history of hepatitis.

These questionable cases, as well as notable abuses of adult research subjects like those in the Tuskegee syphilis study, created the need for stricter regulations in the US. In 1973, the first set of proposed regulations was published, but it did not specifically address children. However, after recognizing the need for pediatric research, particularly with growing evidence that children often do not respond to medications in the same way as adults and suffer from diseases that are unique to childhood, a set of guidelines were created for children. These guidelines were laid out by the National Commission for Protection of Human Subjects of Biomedical and Behavioral Research in 1977, which included the concept of an IRB through which all research involving human subjects must undergo approval. The

Commission acknowledged the need to include pediatric subjects in research studies, but also noted that children represent a particularly vulnerable group. Their vulnerability stems from their inability to provide true informed consent, particularly in younger children, because in order to do so, they must be able to consider the risks and benefits of participating in research for themselves. For this reason, they require additional protections and prohibitions on the kinds of research that can be performed. The main tenets are as follows, “For research with children to be approvable, the research must fit into one of three categories: 1) research not involving greater than minimal risk to participants, 2) research involving greater than minimal risk but with the prospect of direct benefit to individual participants, and 3) research involving no greater than a minor increase over minimal risk (with no prospect of direct benefit) but likely to provide generalizable knowledge of the subject’s condition or disorder that is vital to understand or ameliorate it”.

Ethics of Pediatric Research

Since the publication of the Commission, many ethical questions have arisen, particularly over what is considered minimal risk. Minimal risk is defined “the probability and magnitude of harm or discomfort anticipated in the proposed research are not greater, in and of themselves, than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests.”. This means that research cannot subject children to any more discomfort than they encounter in normal everyday encounters or at a well-child visit. Things that might be approvable would include interviews that do not cause significant stress or anxiety, observation of the child, a single chest x-ray, a single blood draw, a clean-catch or bag urine specimen, or the collection of an additional milliliter of spinal fluid during a clinically indicated lumbar puncture. However, it is important to note that things that adults would consider benign, such as Tanner staging, may be perceived by children to be anxiety inducing. Even interventions that don’t appear invasive, such as questionnaires, have to be given additional thought as they might create anxiety, induce guilt, or encourage certain risky behaviors that could be seen as exceeding minimal risk. Invasive procedures, such as catheterized urine specimens, lumbar punctures, or multiple blood draws are largely seen as going beyond minimal risk. However, nothing is clear cut. In patients with spina bifida for example, who routinely perform self-catheterization to empty their bladders, a catheterized urine specimen might be considered only a minor increase over minimal risk and thus approvable.

Another important ethical consideration for pediatric research is the process of informed consent. Informed consent in children involves a combination of parental (or proxy) permission and/or child assent. Assent is a child’s voluntary agreement to participate in research and is needed in cases where the IRB feels that a child is capable of providing assent based on their age, maturity, and psychological state. Assent does not require a child to be able to make complex medical decisions. Rather, it requires a child to be able to understand that the research is not being done for his or her benefit, to understand what will happen to him or her in the study, and to agree or disagree with participating. Assent is not the same as consent. The purpose of assent is not to view children as autonomous decision makers, akin to adults. Rather, it emphasizes treating them with dignity and respect.

Current State of Pediatric Research in the US

With all these regulations on pediatric research, it comes as no surprise that designing appropriate studies and recruiting children for these studies can be a challenge. Recognizing the dearth of pediatric studies, especially in comparison to the adult population, the FDA came out with the Pediatric Research Equity Act (PREA) and the Best Pharmaceuticals for Children Act (BPCA) in 2012. These helped to increase the number of pediatric clinical trials in a short period of time. In the 5 years following the passage of these acts, 436 studies were done, and roughly 56,000 children were enrolled.

As the number of pediatric studies continues to rise, so has the funding. Currently, the NIH spends roughly ~\$3.6 billion annually for pediatric research. While this is only a portion of the \$41.7 billion spent annually by the NIH, funding will continue to grow as an increased emphasis is placed on the need for quality pediatric research trials. With so many studies in the pipeline, and more continuing to come, the future of pediatric research is looking bright.

References:

1. Bavdekar S. B. (2013). Pediatric clinical trials. *Perspectives in clinical research*, 4(1), 89–99. <https://doi.org/10.4103/2229-3485.106403>
2. Connor, E. M., Smoyer, W. E., Davis, J. M., Zajicek, A., Ulrich, L., Purucker, M., & Hirschfeld, S. (2014). Meeting the demand for pediatric clinical trials. *Science translational medicine*, 6(227), 227fs11. <https://doi.org/10.1126/scitranslmed.3008043>
3. Diekema DS. Conducting ethical research in pediatrics: a brief historical overview and review of pediatric regulations. *J Pediatr*. 2006;149(1 Suppl):S3-11.
4. M.A. Grodin. Historical origins of the Nuremberg Code. G.J. Annas, M.A. Grodin (Eds.), *The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation*, Oxford University Press, New York: (1995), pp. 121-144
5. US Department of Health and Human Services. Federal policy for the protection of human subjects: Additional protections for children involved as subjects in research. (45 CFR 46, Subpart D) *Federal Register*, 48 (1983), pp. 9818-9820

V.14 Evidence Based Medicine in Pediatrics: Unique Challenges and Tools to Overcome Them – (Sarah Banerji, GSM4)

The field of Pediatrics poses unique challenges in the application of Evidence Based Medicine. As Dr. Hira Haq notes in a previous chapter of this book, certain historical factors limited the number of pediatric-specific studies to guide clinical practice. In recent years more pediatricians have called attention to these gaps in research and highlighted the need for increased high quality, evidence-based studies. In fact, the American Academy of Pediatrics (AAP) has put forth multiple policy statements calling for more pediatric research across all spectrums: basic science, translational, community based, health services and child health policy.¹

There are several reasons as to why child health requires unique research. First, childhood affects health throughout life and the long-term effects of interventions in childhood need to be considered. Second, in comparison with adults, children have relatively fewer chronic diseases and more rare diseases that might not otherwise be studied. Third, the physiology of pediatric disease, even chronic diseases common to both adults and kids such as asthma, obesity, and ADHD, is not always the same as adults, nor is the target of treatment.

Despite the importance, there are a few reasons cited for the difficulties conducting research in Pediatrics. These include lack of numbers for rare but concerning diseases as well as therapeutic orphaning from pharmaceutical companies, unique ethics of research in Pediatrics related to consent and parental/guardian roles, retaining children throughout study time, and duration of effects over a long life. Medical advances are often developed for and tested initially in adults, leaving pediatric practitioners with little or no evidence-based guidance on appropriate use in children¹. While the number of high-quality evidence-based studies has increased, there is not always a study to match every question in clinical practice. Still, it is important to continue the search for evidence of benefit and harm in the service of our patients. Often, the practice of Evidence Based Medicine, may reveal a lack of evidence.⁴ In fact, as Dr. Jacobson, previous chair of the Department of Pediatric and Adolescent Medicine at the Mayo Clinic, argues, it is in those moments of seeking answers and finding limited evidence that the quest for investigation is driven. Further, if in fact an extensive search proves that there is uncertainty, this lack of evidence can guide medical decision-making as well.⁴

Thus, while increased pediatric specific research is certainly needed, it is important to develop a framework to assess applicability of studies conducted in non-pediatric patient populations for cases when no good evidence is available. Dr. Bob Philips, who collaborates with the Centre for Evidence-based Medicine in Oxford, UK and the Centre for Evidence-based Child Health in London, UK outlines such a modified framework for assessing clinical studies for pediatricians. He comments that in Pediatrics the first question to consider is whether there are biological differences between the population being studied and the one you are considering treating.⁵ Here it is important to think about

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

the pathogenesis in children and whether the cause of pathology is similar in both populations. Second, he advocates to consider whether “differences in psychology, social setting or economy will stop the data being applicable.” Here it is important to consider whether these factors may influence a family's adherence to therapy. Third, it is important to address issues of risk and co-morbidity and how they might differ from the population originally studied. For instance, if a drug is known to increase the risk of GI bleeding in adults, the pediatrician needs to consider baseline risk of GI bleeding in the age group in question to properly estimate the benefit of the drug. Finally, outcomes must be considered differently in the pediatric population as long-term outcomes of treatment may have unintended consequences for a pediatric patient. With these tools, examining the “biological and psychological differences, consider[ing] the inherent risk and co-morbidities, and examin[ing] all the outcomes closely,” a pediatrician can think more critically about how to apply best evidence to practice.⁵

Despite the historical limitations in conducting pediatric research, in my exploration of the pediatric literature, I was able to find numerous examples of large scale, high-quality, pediatric specific studies, suggesting an increased focus in improving pediatric research since the aforementioned opinions were put forth. This included a large-scale trial demonstrating that prophylaxis with a single dose of Nirsevimab, a monoclonal RSV fusion protein antibody, significantly reduced cases of medically attended RSV in term and near-term infants. The NNT to prevent a case of RSV was 12 (CI 20 to 8)³. Another trial examined the relationship between early food exposure and allergy, demonstrating that the introduction of common allergens in food, namely peanut, cow's milk, wheat, and egg from 3 months of age, complementary to regular feeding, can significantly reduce food allergy at 36 months of age (NNT of 67 with CI 196 to 36)⁶. A third study evaluated whether continuation of the SSRI fluoxetine after initial 12-week therapy in adolescents with Major Depressive Disorder for an additional six months prevented relapse. Strong evidence was found that continuation significantly reduced full relapse in adolescents with a NNT of 3 (CI 18 to 2)². There is certainly room for increased research in Pediatrics, however the quality of clinically relevant research in Pediatrics is growing, making skills of EBM all the more crucial.

The field of Pediatrics has unique challenges regarding the practice of EBM. The increased number of high quality pediatric-specific studies as well as a framework developed to approach suboptimal data highlight the importance of continuing to teach and practice EBM in pediatrics. With a greater emphasis on research, increased funding, better skills of appraisal, and a scholarly approach, Pediatricians will be better prepared to practice and deliver evidence-based care.

References:

1. Cabana MD, Cheng TL, Bauer AJ, Bogue CW, Chien AT, Dean JM, Angela Kelle BS; Promoting Education, Mentorship, and Support for Pediatric Research. *Pediatrics*. 2014 May;133(5):943–949. doi: 10.1542/peds.2014-0448.
- 2.
3. Emslie et al. Fluoxetine versus placebo in preventing relapse of major depression in children and adolescents. *Am J Psychiatry*. 2008 Apr;165(4):459-67. doi: 10.1176/appi.ajp.2007.07091453.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

4. Hammitt, LL et al. Nirsevimab for Prevention of RSV in Healthy Late-Preterm and Term Infants. *N Engl J Med.* 2022 Mar 3;386(9):837-846. doi: 10.1056/NEJMoa2110275.
5. Jacobson RM; Association of Medical School Pediatric Department Chairs, Inc. Pediatrics and evidence-based medicine revisited. *J Pediatr.* 2007 Apr;150(4):325-6. doi: 10.1016/j.jpeds.2006.12.044. PMID: 17382101.
6. Phillips B. Towards evidence based medicine for paediatricians. *Arch Dis Child.* 2004 Mar;89(3):286-7. doi: 10.1136/adc.2003.048280. PMID: 14977717.
7. Skjerven HO et al. Early food intervention and skin emollients to prevent food allergy in young children (PreventADALL): a factorial, multicentre, cluster-randomised trial. *Lancet.* 2022 Jun 25;399(10344):2398-2411. doi: 10.1016/S0140-6736(22)00687-0.

Submitted 12/9/2022

V.15 Applying Population-based Studies to the Individual Patient (Fatima Haidar, GSM 4)

Overview

The results of a high-quality, practice-changing Randomized Control Trial (RCT) have just been published. As a physician, you pride yourself in practicing Evidence-Based Medicine (EBM), but you now face a challenge: How will you incorporate the results of this population-based trial into the management of your *individual* patient? In the following sections, we will explore the methods and some statistical exercises appropriate for answering this question.

Questions on Applicability

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

RCTs are firmly established as the gold standard of EBM, in part because they are the only study design capable of proving causality between interventions and outcomes¹. However, there are a number of important factors to consider prior to applying the results of a population-based RCT to an individual patient². Below are four questions, adapted from Glasziou et al., that provide a framework to consider whether a population-based study applies to your patient:

1. *Is my patient similar to the study participants?* To answer this, review whether your patient meets the inclusion criteria of the study. If your patient does meet the inclusion criteria, then ensure that the baseline characteristics of the study participants are similar to those of your patient. If they are, then proceed to the next question. If they are not, then consider whether the baseline characteristics between study participants and your patient are different enough to limit the applicability of the study results.
2. *Is this treatment similarly accessible for my patient?* There are a number of factors to consider when answering this question, including your patient's Social Determinants of Health and the limits of your practice setting. Such factors may impact the applicability of the study results to your patient. For example, if a drug is reported to carry a 3% risk of mortality due to hemorrhage, will this risk of mortality be higher in your patient who lives 2 hours away from the nearest ED? If the treatment is similarly accessible for your patient, then proceed to the next question.
3. *How will my patient's autonomy influence this decision?* Shared decision-making is crucial to the successful management of patients. Our patient's goals of care, beliefs, and values should be solicited prior to suggesting treatment options. In addition, if our patient lacks capacity, then it is important to seek information on goals of care and treatment preferences from official documents, such as an Advanced Directive, or from family members and friends who knew the patient prior to loss of capacity. If you agree that a treatment likely aligns with your patient's goals of care, then proceed to the next question.
4. *What are the benefits and harms to my patient?* Learning how to answer this fourth and final question is the goal of the remainder of this chapter. By the end of this chapter, you should be able to individualize the risks and benefits of a population-based study to your individual patient.

Statistics of Individual Benefit and Harm

The results of an RCT are, by nature, population-based. Outcomes such as Relative Risk Reduction (RRR), Absolute Risk Reduction (ARR), and Number Needed to Treat (NNT) are all derived from – and thus, applicable to – populations. But what do these population-based outcomes reported in an RCT mean for the individual patient? A number of methods exist for individualizing the population-based outcomes reported in RCT. In this section, we will start by learning how to individualize the NNT.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

The NNT is defined as the number of patients that we would expect to be treated before one achieves the desired benefit of the treatment. For example, if Drug X has a NNT of 10, that suggests that for every 10 patients treated with Drug X, we would expect 1 patient to achieve the desired benefit of Drug X. However, the patients enrolled in an RCT may be quite different from the patient in front of you. How can you individualize the NNT to be more useful to your patient?

To do this, we can calculate the **patient-specific NNT**. The patient-specific NNT combines the RRR (reported by an RCT) and the **patient-expected event rate (PEER)**³. The PEER is defined as the likelihood of an adverse event occurring in your patient before treatment. We will use the following formula to calculate the patient-specific NNT:

$$\text{Patient – specific NNT} = \frac{1}{(\text{PEER} \times \text{RRR})}$$

Now, let's practice calculating the patient-specific NNT using the case of Mr. T, who is a 76 year old man with dyslipidemia, type 2 diabetes, and hypertension. He does not currently take any medications. He does not smoke or drink. His 10-year risk of death from heart disease or stroke (i.e. his PEER) is 36.7%, based on the ASCVD risk calculator⁵.

Luckily for Mr. T, a recent RCT demonstrated that Drug X may help patients like him reduce their risk of mortality from heart disease or stroke. In this RCT, investigators reported that Drug X has a NNT of 10 and resulted in a 70% RRR in mortality from heart disease or stroke. When we enter these values into the equation above, we obtain the following:

$$\text{Mr. T's NNT} = \frac{1}{(0.367 \times 0.7)} = 4$$



Through this calculation, we find that Mr. T's NNT is 4; however, as you might recall, the NNT reported in the RCT was 10. Why is that? While these results seem conflicting, they suggest that patients like Mr. T – that is, patients who have similarly elevated baseline mortality risk due to heart disease or stroke – are more likely to benefit from Drug X than were the patients who were enrolled in the study. Put more simply, if we have 5 patients with a baseline mortality risk (i.e. PEER) similar to that of Mr. T, and we treat them all with Drug X, then we anticipate 1 of those patients will have a $\geq 70\%$ reduction in mortality risk and the other 4 patients will not (Fig. 2). On the other hand, the NNT of 10 reported in the RCT suggests that we have 11 patients with a baseline mortality risk similar to those in the RCT, and if we treat them all with Drug X, then we anticipate 1 patient will have a $\geq 70\%$ reduction in mortality risk and the other 10 patients will not. In short, the patient-specific NNT suggests that Mr. T is more likely to benefit from Drug X than would be suggested by the NNT reported in the RCT.

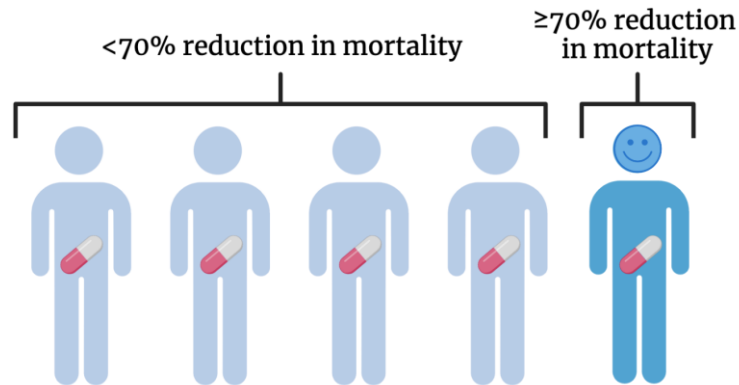


Figure 1. This figure depicts the patient-specific NNT calculated for Mr. T, which was calculated above. The NNT of 4 suggests that for every patient expected to achieve a ≥70% mortality risk reduction on Drug X, there are 4 other patients on Drug X who would not be expected to achieve this result.

While this patient-specific NNT is a useful outcome for understanding how the results of an RCT apply to your individual patient, it is often a difficult concept to understand or to explain. In this case, Mr. T thanks you for explaining his patient-specific NNT, but asks if there is a simpler way to explain how Drug X would benefit him. “After all,” he asks. “What if I’m among the 4 who don’t improve on Drug X?” In fact, as the NNT increases, as is more common with many clinical trials, the reasonableness of asking that question is even more important to address.

Using the outcomes from the RCT above, we can calculate another patient-specific outcome: the **patient-specific risk reduction**. Similar to the patient-specific NNT, the patient-specific RR individualizes the RRR reported in an RCT by incorporating our patient’s PEER. It is calculated using the following formula:

$$\text{Patient – specific mortality risk reduction} = (\text{RRR} \times \text{PEER}) \times 100$$

Again using the case of Mr. T, we will input the RRR reported by the RCT (70%) and Mr. T’s estimated 10-year mortality risk (36.7%) into the formula above, and obtain the following:

$$\text{Patient – specific mortality risk reduction} = (0.7 \times 0.367) \times 100 = 25.7$$

The results of this formula show that Mr. T’s 10-year mortality risk, which was 36.7%, will be reduced by 25.7% with Drug X. Therefore, rather than a 10-year mortality risk of 36.7%, Mr. T will have a 10-year mortality risk of 11%. Upon sharing this statistic with Mr. T, he is thrilled and feels that this reduction in mortality is significant enough to start treatment with Drug X. Strong work, doctor!

The Likelihood of Being Helped or Harmed (LHH)

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

There is one final statistic we should consider: the **Likelihood of Being Helped or Harmed (LHH)**. The LHH is defined as the ratio of the likelihood of benefit to the likelihood of harm⁴. For example, an LHH of 10 suggests that a patient is 10x more likely to benefit from a treatment than to be harmed by it. In brief, the LHH answers the following question: Is my patient more likely to benefit or be harmed by this treatment?

However, it is important to note that there are multiple ways to calculate the LHH, as “benefit” and “harm” are defined by you. That is, you specify what “benefit” and “harm” you are interested in. Examples of benefits include outcomes such as complete response, partial remission, or reduction in mortality, while examples of harms include outcomes such as all-cause discontinuation, adverse events, or increase in mortality.

One common calculation of LHH combines two very useful outcomes – the NNT and the Number Needed to Harm (NNH) – into one outcome. Remember, the LHH is the likelihood of benefit to harm; therefore, NNT:NNH is one way to calculate the LHH. This can be done in one of two ways. The faster way is to calculate the LHH by using the NNT and NNH reported in an RCT. The second way is to calculate the patient-specific LHH using the patient-specific NNT and NNH; this allows us to further individualize the results of an RCT to our patient.

The formula and sample calculations for patient-specific NNT were completed in the previous section; below is the formula for calculating the patient-specific NNH:

$$\text{Patient – specific NNH} = \frac{1}{(\text{PEER} \times \text{RRI})}$$

For example, consider that an RCT reports Drug Y has a NNH of 17 and a Relative Risk Increase (RRI) of 10% in annual stroke risk. You are considering starting Drug Y on your patient, but would like to calculate their patient-specific NNH. Your patient’s CHADsVASC score is 9, suggesting a 17.4% annual risk of stroke, TIA, or systemic embolism⁵. When we enter these values into the formula above, we obtain the following:

$$\text{Patient – specific NNH} = \frac{1}{(0.174 \times 0.1)} = 57$$

In this example, your patient’s NNH is significantly higher than that of the population studied in the RCT. This would suggest that your patient may have an increased risk of harm from Drug Y than did the study participants. Knowing this, you may hesitate to start Drug Y on your patient, who is more likely to experience harm than those who participated in the study.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Lastly, while the LHH is a valuable statistic that combines benefit and harm into one outcome, it is also important to consider the limitations of the LHH. Remember, the LHH is a ratio of two values – often the NNT and NNH – so it provides no information on the significance of the benefit or harm seen. For example, an LHH of 5 may suggest that a patient is 5x more likely to experience benefit than harm, but if the NNT is 5,000 and the NNH is 1,000, then there is still a significant chance of harm; that is, for every 5 people who benefit, 1 is harmed. Therefore, it is important to consider a variety of statistical outcomes when applying the results of a population-based study to your individual patient, as each statistical outcome has its own unique benefits and limitations.

Conclusion

In this chapter, we covered select topics related to the application of population-based studies to individual patients. We reviewed a framework of four questions to gauge the applicability of a population-based study, such as an RCT, to your individual patient. Upon establishing a study's applicability, we learned about a number of statistical outcomes that can be used to individualize the results of a study to your patient. Specifically, we learned how to calculate a patient-specific NNT, RR, NNH, and LHH. With this information, we will be better equipped to understanding if, when, and how the results of population-based studies apply to our individual patients.

References

1. Channer KS. Translating clinical trials into practice. *Lancet*. 1997 Mar 1;349(9052):654. doi: 10.1016/s0140-6736(05)61606-6. PMID: 9057761.
2. Glasziou P, Guyatt GH, Dans AL, Dans LF, Straus S, Sackett DL. Applying the results of trials and systematic reviews to individual patients. *ACP J Club* 1998; 129:A15–16.
3. Shaneyfelt, Terry. "How to Calculate Patient-Specific Estimates of Benefit and Harm from a RCT." *Ebmteacher*, 24 Feb. 2015, <https://ebmteacher.com/2015/02/24/how-to-calculate-patient-specific-estimates-of-benefit-and-harm-from-a-rct/>.
4. Andrade C. Likelihood of Being Helped or Harmed as a Measure of Clinical Outcomes in Psychopharmacology. *J Clin Psychiatry*. 2017 Jan;78(1):e73-e75. doi: 10.4088/JCP.16f11380. PMID: 28129502.
5. "Medical Calculators, Equations, Scores, and Guidelines." *MDCalc*, <https://www.mdcalc.com/>.

Submitted 12-29-2022

V.16 Statistical Incorporation of Patient Preferences and Values (Matt Wesley)

Shared decision making is used to determine a patient's underlying preferences and values as it relates to a particular medical decision. Providers contribute an understanding of medical knowledge and communicate the risks and benefits for a given decision. There are detailed methods for assigning patient-specific numerical values to all possible outcomes and decisions through a Clinical Decision Analysis (CDA). However, the time investment and complexity are generally prohibitive for regular clinical use. In a more concise manner, risks and benefits can be communicated as the number needed to treat (**NNT**) and number needed to harm (**NNH**). Combined, the ratio is called the Likelihood of being Helped and Harmed (**LHH**).

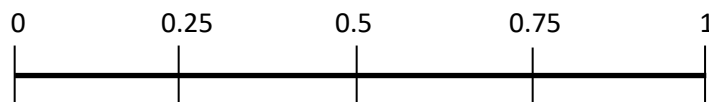
Let's take a hypothetical example of starting a high intensity statin for secondary prevention of cardiovascular disease. If 50 patients need to be treated to prevent one death (**NNT**) and 20 patients need to be treated for one patient to experience myalgias (**NNH**), then the LHH is 0.4. Another way to express this, for all patients in the study population, is that the intervention is 2.5 times more likely to harm than to help.

$$\text{LHH} = (1 / \text{NNT}) / (1 / \text{NNH}) = (1/50) / (1/20) = 2/5$$

The LHH provides patients with a generic estimate of the impact for a given decision. This ratio does not consider patient demographics or patient preferences or values. A patient expected event rate (PEER) can be estimated if a study has a large enough sample size to report subgroup analysis relevant to the patient.

$$\text{NNT}_1 = 1 / (\text{RRR} \times \text{PEER})$$

However, if this information is not available, more specific estimations of the LHH can still be achieved by asking the patient to assign utility to **safety** (a known complication) and **target** (clinical endpoint if left untreated) outcomes. This can be done with a visual aid.



Adapted from Straus' Evidence Based Medicine: How to Practice and Teach It

The patient is asked to separately rate the safety and target outcomes on this scale, comparing "0" to the worst possible state and "1" to the best possible state. Let's use an example patient deciding to undergo laminectomy for cervical spinal stenosis. Assume an NNH of 1000 for the safety outcome of paralysis from surgery and an NNT of 2 for the target outcome of chronic pain. When asked to rate the safety outcome of paralysis from surgery, the patient chooses "0.01" when compared to a "0" representing a vegetative state. When asked to rate the target outcome of

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

chronic pain, the patient values “1” as full ADLs/function and rates untreated chronic pain “0.7” accordingly.

$$\text{LHH} = (1 / \text{NNT}) / (1 / \text{NNH}) = (1 / 2) / (1 / 1000) = 500$$

For a generic patient, this procedure would be 500 times more likely to help than to harm. The utility of the patient’s preferences can be introduced into the LHH calculation as a ratio.

$$\frac{[1 - U_{\text{target}}] / [1 - U_{\text{safety}}]}{\text{LHH}_1 = ([1 / \text{NNT}] \times [1 - U_{\text{target}}]) / ([1 / \text{NNH}] \times [1 - U_{\text{safety}}])}$$

U_{target} is the target utility (0.5 in this example), and U_{safety} is the safety utility (0.01). This patient interprets the safety outcome as very close to the worst possible health state. We find that the calculus changes but still favors the intervention by a factor of 150.

$$\text{LHH}_1 = ([1 / 2] \times [1 - 0.7]) / ([1 / 1000] \times [1 - 0.01]) = 150$$

This example shows despite rare but catastrophic safety outcomes, the target outcome utility for a given patient has a more significant impact on the LHH. The more attractive the target outcome (e.g., approaching “1” on the outcome scale), the more favorable the intervention.

References:

1. Straus SE, McAlister F. Applying the results of trials and systematic reviews to our individual patients.
2. *Evidence-Based Mental Health* 2001;4:6-7.
3. Straus, Sharon E., et al. *Evidence-Based Medicine: How to Practice and Teach It*. Churchill Livingstone, 2011.
4. Straus, Sharon E. "Individualizing treatment decisions: the likelihood of being helped or harmed." *Evaluation & the health professions* 25.2 (2002): 210-224.

September 2018

V.17 Minimal Clinically Important Difference – (Meg Hanley, GSM4)

How do we capture outcomes that are important to our patients?

Patients seek out interventions – medicines, therapies, surgeries – to feel better. Physicians study interventions to understand their safety, efficacy and typically to compare the interventions to others. As a research study progresses through sequential phases from feasibility through safety to efficacy, it becomes increasingly important to consider not just treatment effect but clinical outcome(s). Patient reported outcomes are increasingly a part of clinical trial design; the Patient-Centered Outcomes Research Institute (PCORI) exists specifically to fund clinical effectiveness research to furnish actionable information for patients and providers.

Minimal clinically important difference (MCID) is an approach to understanding clinical significance that is both patient centered and enables interpretability and applicability across studies. Alternative related terms include minimally important difference (MID) and minimal clinically important improvement (MCII). MCID seeks to define the smallest amount of change in outcome that *patients* deem important. Put best by those who coined it, “minimal clinically important difference can be defined as the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management” (Jaeschke).

Just as a p-value can be used to determine *statistical* significance, achievement of a MCID provides insight into *clinical* significance. Since p-value is tied to sample size, MCID becomes increasingly important as the number of study participants increases because statistical significance may occur in a large sample even with only small differences which are often clinically meaningless.

MCID enables a calculation of % of patients improved by the intervention – a digestible takeaway for clinicians regardless of the amount of time they have to dedicate to reading papers. An alternative metric in use is the number needed to treat (NNT) calculated as $1/\text{absolute risk reduction}$, as a way to say “X number of patients need to be treated to prevent this adverse outcome”. However this does not capture what the patient perceives or values.

How do we determine MCID for a study population?

There are three commonly employed methodologies for determining MCID: consensus, anchor and distribution-based. The *consensus (or Delphi) approach* relies on an expert panel to define independently a clinically relevant change. Each panelist provides an assessment, then the panelists review each other's assessments with iterative revisions until a consensus, a singular MCID value, is reached. Of note, experts are not patients. A major critique of the consensus approach, therefore, is

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

that experts in a specific domain may not accurately perceive the degree of change which matters to a patient. Experts may be motivated to report on lesser change as clinically important in order to preserve meaningful results. On the flipside, experts have interacted with the exact patient population in question and have honed a clinician “gut-based” expertise from years of listening to patient reported outcomes. The Delphi approach was put to use by an expert panel of 6 rheumatologists exploring the effect of NSAIDs in osteoarthritis. In the first round, each expert reviewed a provided study with data set and came to an individual determination of MCID for 41 unique outcomes. The six resulting MCIDs were tabulated anonymously and in the second round the 6 experts were provided with the same data and the table of recommended MCIDs and were invited to modify the MCID they proposed. A third round followed, identical to the second with experts given a final opportunity to modify their numbers. At no point was rationale or calculation shared between experts. Overall, the range of MCIDs decreased for all but one outcome from round 1 to round 3 (Bellamy).

The *anchor approach* is based on patients’ qualitative assessments of their personal response to an intervention. This method relies on patient-facing surveys with qualitative descriptions of change that are tied to a quantitative scale, thereby anchoring the measurable change to a descriptive change. In anchoring, the choice of anchor (the subjective assessment) is key. Different anchors vary in validity and the choice of anchor may be subject to specific biases. For example, asking a patient about their improvement in symptoms may lead to recall bias. Additionally, data outliers can skew anchoring data, but deriving MCID from only a subset population ignores the vast inter-patient variability. For example, Tubach et al. used an anchoring method to determine MCID in osteoarthritic pain in patients who trialed NSAIDs. To do so, 1362 patients with osteoarthritis were studied during a course of treatment with NSAIDs. Pain, on a 0-100 mm VAS scale was assessed at baseline and final visit. Additionally, at the final visit two-thirds of patients completed surveys on their perceived response to NSAID treatment, using a five point Likert scale: none - no good at all; poor - some effect but unsatisfactory; fair - reasonable effect but could be better; good - satisfactory effect with occasional episodes of pain or stiffness; excellent - ideal response, virtually pain free. MCID in pain was calculated based on absolute (final value-baseline) and relative (final value-baseline/baseline) changes in each of the three patient reported outcomes (see Tubach, 2005 for statistical approach).

Tubach et al. found that an absolute decrease in pain of 19.9 points (relative decrease of 40.8%) was the minimal clinical important difference for NSAIDs use in the OA population. Since then, multiple studies have continued to use the MCID Tubach and colleagues derived for pain in OA (Kuebler 2022, French 2022).

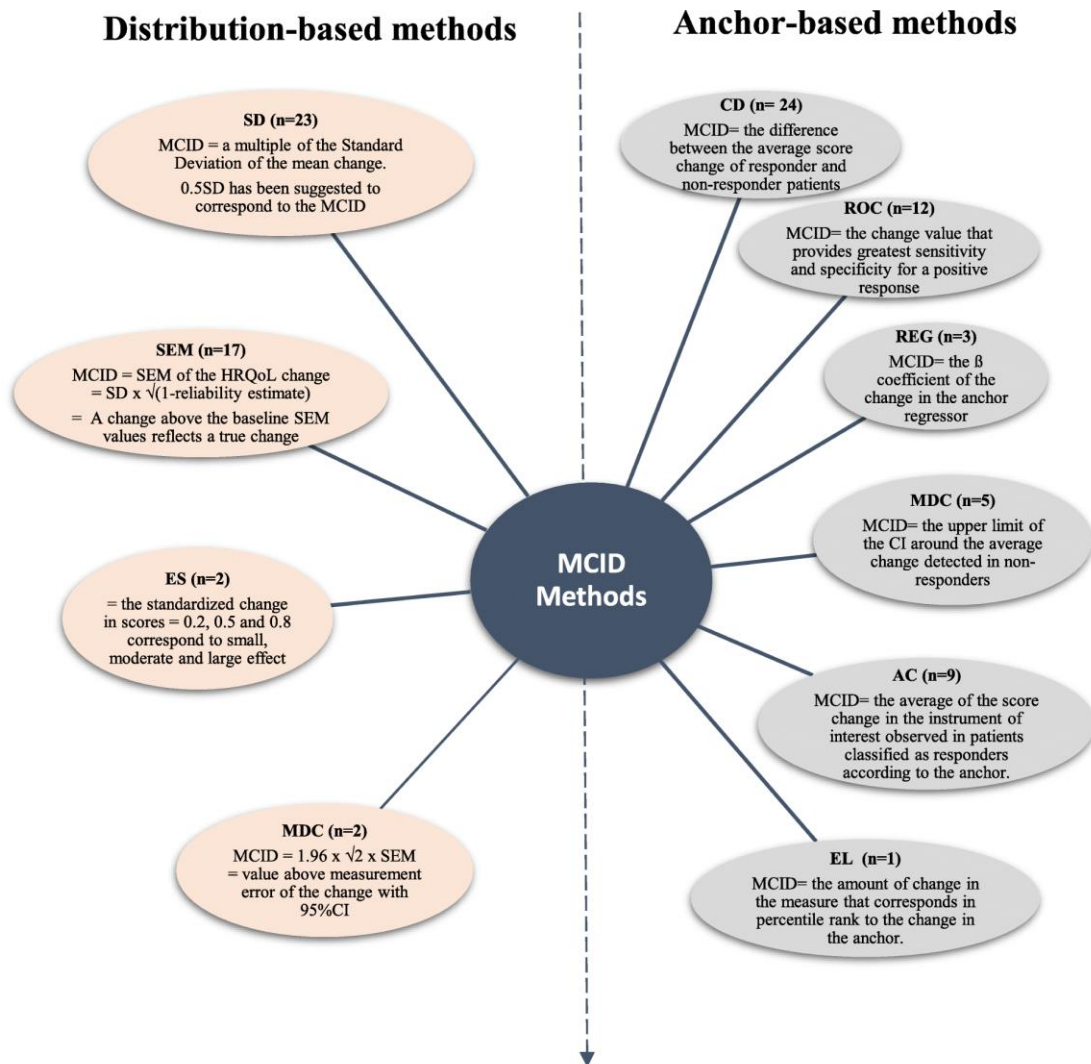
The *distribution-based approach* to determining clinical importance of an outcome is based solely on statistics: standard deviation, standard error and effect size. As such, it is best described as the minimal detectable effect; and effect that is unlikely to be due to random error (McGlothlin, 2014). The distribution of the study’s scores and the variation between scores is the basis of this approach. A benefit of this approach is that it does not require an expert panel to convene and does not require use of patient assessments like survey. The latter, lack of tie to patient assessment, is also the Achilles heel of the distribution-based approach; patient experience/opinion does not factor into the

determination of important difference. As McGlothin et al concludes, a non-patient facing, purely statistical approach is not an appropriate method for determining clinical import.

I'm studying X, how do I include MCID?

If you are studying an intervention with effects on patients, then there is likely room to include MCID for an outcome of study. Remember, MCID is not only limited to improvements but can also be used to quantify the negative effects of an intervention that are important to a patient including side effects of a study drug or cost of a therapeutic intervention. The first step to include MCID is to investigate whether MCID has been established for your population of study and intervention.

In a systematic review of the methodology of MCID determination for quality-of-life instruments, 47 studies showed great variation in how MCID can be calculated, as shown below (Moelhi 2020). Anchoring based approaches were most common, either alone or in combination with distribution-based methods. The most used anchoring method was "change difference" (CD), defined as the mean change of patients who improved anchored to cutoffs based on a scored degree of change, such as "responder" or "non-responder". In the visual below, the statistical nature of distribution-based methods vs. the descriptive, patient response focused nature of the anchoring method is clear.



Note: more than one statistical method was used in most studies.

MCID: Minimal Clinically Important Difference, AC: Average Change, MDC: Minimal Detectable Change, CD: Change Difference, ROC: Receiver Operating Curve, REG: Regression analysis, EL: Equipercentile Linking, SD: Standard deviation, SEM: Standard error of measurement, ES: Effect size

What would the ideal future of MCID look like?

In an idealized world, we would have an established MCID for all common outcomes of study for every disease group. In addition, we'd have MCIDs for relevant subsets of study population such as age, gender, or baseline as applicable to outcome. Consider that MCID may be different for subset populations: an incontinent patient who requires 10 pads per day, may only see improvement in quality of life if their symptoms improve to warrant 5 pads a day, whereas a patient using 4 pads per day may experience a two-pad reduction as important. In such a world, MCID for subsets of populations would

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

enhance provider-patient shared decision making. Lastly, if we had an MCID for all interventions of common disease, there would be an additional metric for comparison across research and providers. In the absence of an idealized world, there may be value in using multiple methods – anchoring and consensus or anchoring and distribution to “triangulate” towards an MCID for a given intervention. Additionally, a responder-only analysis in which those patients who achieved MCID of an outcome are included could yield enhanced understanding of response. By removing outlier data in responder-only analysis, the results may advance personalized medicine and better our understanding of who can be reasonably expected to respond to the intervention.

References

Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis.* 1987;40(2):171-178. doi:10.1016/0021-9681(87)90069-5

McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA.* 2014;312(13):1342-1343. doi:10.1001/jama.2014.13128

Bellamy N, Carette S, Ford PM, et al. Osteoarthritis antirheumatic drug trials. III. Setting the delta for clinical trials--results of a consensus development (Delphi) exercise. *J Rheumatol.* 1992;19(3):451-457.

Tubach F, Ravaut P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. *Ann Rheum Dis.* 2005;64(1):29-33. doi:10.1136/ard.2004.022905

Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10(4):407-415. doi:10.1016/0197-2456(89)90005-6

Kuebler D, Schnee A, Moore L, et al. Short-Term Efficacy of Using a Novel Low-Volume Bone Marrow Aspiration Technique to Treat Knee Osteoarthritis: A Retrospective Cohort Study. *Stem Cells Int.* 2022;2022:5394441. Published 2022 Nov 15. doi:10.1155/2022/5394441

French HP, Abbott JH, Galvin R. Adjunctive therapies in addition to land-based exercise therapy for osteoarthritis of the hip or knee. *Cochrane Database Syst Rev.* 2022;10(10):CD011915. Published 2022 Oct 17. doi:10.1002/14651858.CD011915.pub2

Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR; Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc.* 2002;77(4):371-383. doi:10.4065/77.4.371

Mouelhi Y, Jouve E, Castelli C, Gentile S. How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. *Health Qual Life Outcomes.* 2020;18(1):136. Published 2020 May 12. doi:10.1186/s12955-020-01344-w

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Submitted 12/11/2022

V.18 Decision Aids and Shared Decision Making (George S. Wang and Jesus Mendez Jr, GSM4)

Clinicians must be able to provide patients with the best available information and identify what individual patients value. Shared decision-making (SDM) has been widely accepted as a critical feature in high value care. However, the way that clinician presents information to patients can strongly influence their decision-making. Decision aids are a methodology that enables the information to be provided in an unbiased manner and the patient to introduce their own values into the process. To best help patients, the decision aid has to take into account the strength of the evidence but also they need to translate the probabilistic nature of the evidence for individual patients to help them reach their decision based on informed values. There have been many studies looking into the benefits and outcomes of using different decision aids. This review will walk through the different aspects of decision aids and when they can be useful.

There are a multitude of different decision aids aimed at helping to support individual medical decisions. The goal of these decision aids is to supplement rather than replace physician counseling about options. However, in general these decision aids will follow the following principles:

1. Explicitly state the decision that needs to be considered
2. Provide evidence-based information about the options, benefits, harms, and probabilities of each.
3. Help patients determine the value sensitive nature of their decision and help them clarify the values they placed on the individual benefits and harms.

There is still concern about the large variability in the quality of patient-oriented information. Three domains of quality have been determined to be helpful: clinical content, development process, and the evaluation of a patient's decision aid's effectiveness. However, standards and certifying criteria are still being actively developed for patient aids [1].

The decision aid can be used at multiple points during a patient's time in a clinical encounter. Giving the patient that decision made before the clinical encounter allows patients ample time to research the topic and come prepared to ask questions during the encounter. Providing the decision aid during the clinical encounter can allow the physician to walk the patient through the process. Finally providing the decision made after the encounter can give the patient time to make the decision before a future clinical encounter. When to present the decision aid should be tailored to each individual physician's practice and can be trialed to find the best fit.

There has been evidence published that supports the important role that decision aids can make in a patient's clinical encounter. A 2017 systematic review found that patient decision aids were helpful in multiple ways. Patients had increased knowledge, and better perception regarding the risks associated with either choice from the decision they had to make compared to the usual care [2].

Implementing in practice

The following is an example of a common scenario where a decision aid could be implemented:

Case: Dr. T is a primary care physician of Mrs. D, a 45-year-old woman who presents today with questions regarding breast cancer screening. Her friend was recently diagnosed and she wants to ensure that she does not have breast cancer as well. Dr. T realizes that he has many patients in a similar situation as Mrs. D who come to him with questions regarding breast cancer screening and whether or not they should undergo it. He decides to look into a decision aid to help patients better understand the benefits and harms of screening.

1. **Identify the decisions involved.** For this to be effective, both the patient's point of view and the provider's point of view has to be taken into account. At this point it may also be necessary to conduct a survey of the patient's needs and/or review of the literature regarding this particular decision.

Dr. T believes that to undergo the screening, there could be risk of overdiagnosis and subsequent overtreatment of breast cancer. However, there is a chance that Mrs. D does have underlying breast cancer which would benefit from early diagnosis. Mrs. D has the option of undergoing screening or waiting until a later age.

2. **Create the decision aid.** Make sure to include a way for patients to incorporate their own values. The Ottawa hospital research institute (www.ohri.ca) is a good place to start; the decision aid library inventory, and med-decs can also be used as resources for guides and a constantly updated library of patient decision aids [2].

*Dr. T does a little bit of research and finds a breast cancer screening decision aid made on the Ottawa hospital research institute library of decision aids site:
<https://www.healthdecision.org/tool#/tool/mammo>.*

3. **Identify barriers to implementation.** Ask patients and providers what barriers may exist to providing the patient decision aid. Is the problem with the decision aid? Provider? Or current patient population? Finding these barriers before implementation before the decision aid is implemented can greatly ease the process.

Dr. T talks to his staff to find when would be the best time to implement the decision aid. They come to the agreement that after a patient expresses the desire or questions whether they may need to undergo breast cancer screening, your decision aid should be sent home with the patient and the follow-up visit can be scheduled.

4. **Implementation.** Synthesize the information gained from steps 1 through 3. It's important to emphasize the gap between the patient's decision-making needs and the current practice. It is also important to provide training for usage of the patient decision aid tool to all the providers who will be making use of it.

Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine

Mrs. D is sent home with the patient decision aid. She returns with the following results:

Instructions **Data** Assessment Decision Library Patient Info Chart Note Credits **More** ▾

> **Do you have high-risk features?**

- > New breast symptoms?
- > Past breast cancer?
- > Past chest radiation?
- > Known genetic markers?
- > Refuse Cancer Treatment?

"No" for all high-risks ->

These conditions place patients at higher than average risk, or otherwise reduce the value of screening.
 The tool's calculations are designed for average risk women and are less accurate for higher risk women.


Your Data:

- *Patient's Age (35-74) INFO
- Race / Ethnicity INFO
- Previous Breast Biopsy INFO
- Family History - 1st deg. relative INFO
- Breast Density INFO

Continue

Instructions Data **Assessment** Decision Library Patient Info Chart Note Credits **More** ▾

Your Risk

2.3% risk of breast cancer in next 10 years. INFO 
 The average risk for a 45 year old White woman is: **2.7%** [Show Bar Graph](#)

Guideline Recommendations - Selected for this patient

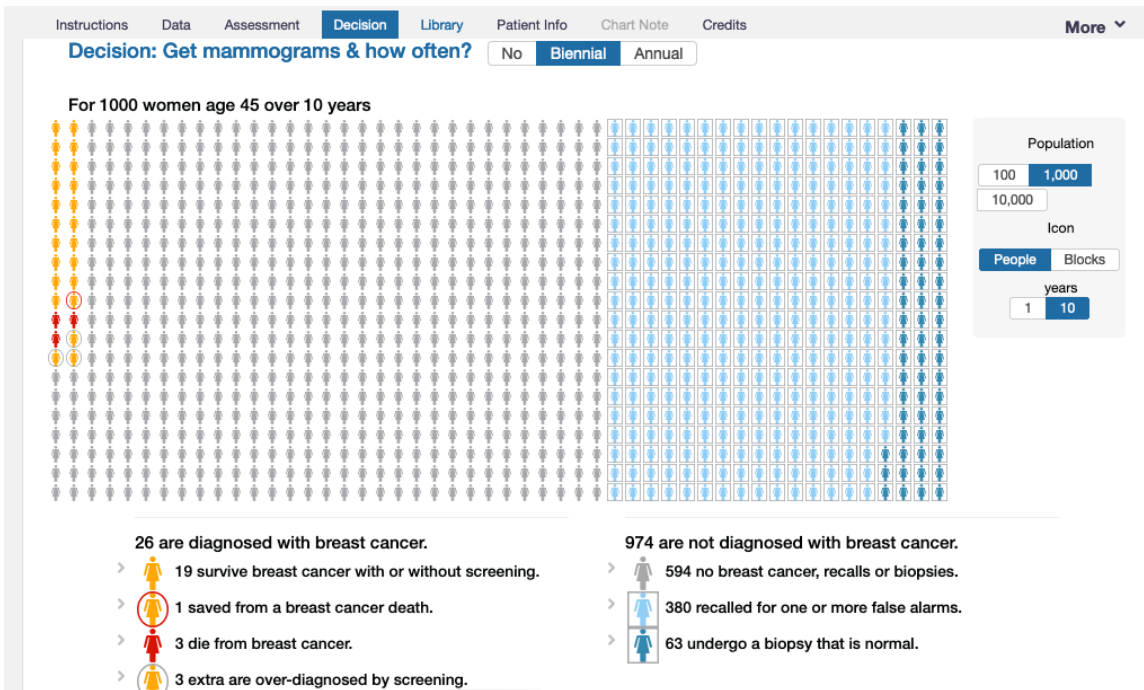
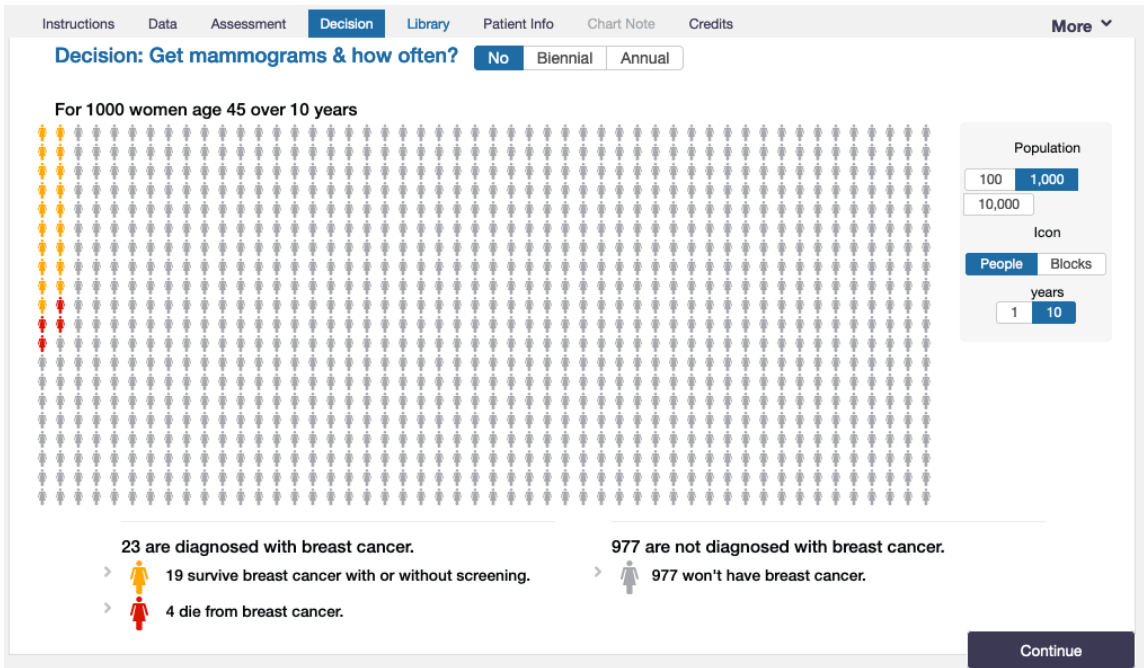
- > Get mammograms every 2 years or wait until age 50. (USPSTF 2016) INFO
- > Get a mammogram every year. (American Cancer Society 2015) INFO
- > Get mammograms every 1 or 2 years or wait until age 50. (ACOG 2017) INFO
- > Get a mammogram every year. (American College of Radiology 2017) INFO

The Choice: Get mammograms or not and how often? (Next page compares options)

> Shared decision-making points:

Continue

Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine



Dr. T addresses Mrs. D's questions and she decides that the risk of having breast cancer is too small and the chance of being over-diagnosed or being called back for an unnecessary biopsy is too much of a hassle for her. She decides to postpone breast cancer screening until she is 50 or starts having changes in her current status. She thanks Dr. T for the helpful tool.

5. Monitor use and outcomes.

It is important to monitor how many providers are making use of the decision aid and how the decision aid is being used. are there any barriers to the usage of the decision aid? How have the decision aids affected the patients? Are our patients making higher quality decisions based on better information? Providers can measure how comfortable patients are now through decision using "the sure test" which is a four-item test which assesses whether the patient feels informed about their options and adequately supported about their decision.

Dr. T continues to use the decision aid over the next few years and asks patients to fill out a sure test survey following but usage of the decision aid. Most patients have a score of 4 on the sure test. Dr. T is very happy to have implemented this decision aid and his practice.

SURE Test version for clinical practice

Yes equals 1 point

No equals 0 point

If the total score is less than 4, it indicates the probability that a patient experiences clinically significant decisional conflict.

		Yes [1]	No [0]
Sure of myself	Do you feel SURE about the best choice for you?		
Understand information	Do you know the benefits and risks of each option?		
Risk-benefit ratio	Are you clear about which benefits and risks matter most to you?		
Encouragement	Do you have enough support and advice to make a choice?		

Decision aids can be useful to clinicians who routinely put difficult screening or therapy decisions before patients. Patients benefit from better knowledge and a sense that they had more control over the decision-making process. But for these tools to be effective, several conditions may be necessary for successful implementation, including good quality decision aids that meet the needs of the population; clinicians who are willing to use decision aids in their practice; effective systems for delivering decision support; and clinicians and healthcare consumers who are skilled in shared decision making.

Another example

I want to engage in shared decision making but I'm not sure how to best convey the risks and benefits for the interventions available for my patient. Are there are any tools I could use?

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Luckily there are a variety of decision aids available to providers to help provide important information to patients and take into account their preferences for which intervention to choose. Decision aids can be useful in increasing patient knowledge on their condition and the available treatment and management options as well as encourage patients to have more input into medical decision making. They can vary in complexity from a simple infographic to full presentations to interactive web tools. There are a wide variety of pre-made decision aids for a wide variety of medical conditions and interventions from starting acne therapy to deciding if weight loss surgery is right for a patient.

The Ottawa Health Research Institute maintains a list of current decision aids that is freely available at <https://decisionaid.ohri.ca/AZlist.html>.

How do I know if the decision aid is effective and up to date?

The International Patient Decision Aid Standard collaboration (IPDAS) regularly review decision aids using a checklist assessing various aspects including whether it provides useful up to date information in an easy to understand and unbiased manner and includes a method for patients to incorporate their values into the process. OHRI uses the IPDAS framework and provides a summary (including it's IPDAS score and it's most recent update) for each decision aid prior to linking you to the actual decision aid.

The checklist used by IPDAS can be found at http://ipdas.ohri.ca/IPDAS_checklist.pdf.

What's an example?

Let's say a patient is considering starting on a daily aspirin to reduce the risk of heart attack. Looking at the OHRI list of decision aids, there appear to be two available and randomly picking one of them will lead to a webpage describing the target audience, the options it includes, last update, format, and who made the decision aid. One might choose <https://statindecisionaid.mayoclinic.org/> made by the Mayo Clinic.

Where can I learn more about shared decision making?

Here's a list of resources:

<https://decisionaid.ohri.ca/>

https://med.dartmouth-hitchcock.org/csdm_toolkits.html

<https://shareddecisions.mayoclinic.org/>

This simple decision aid comes with a surprising amount of information in a single page. It simply covers the main benefit and main harm of taking aspirin (reducing risk of heart attack and bleeding) and shows a visual depiction of the risks and benefits allowing the patient to make an informed choice.

References:

1. Elwyn G, O'Connor A, Stacey D, Volk R, Edwards A, Coulter A, et al. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. *BMJ* 2006

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

2. Stacey D, Légaré F, Lewis K, Barry MJ, Bennett CL, Eden KB, Holmes-Rovner M, Llewellyn-Thomas H, Lyddiatt A, Thomson R, Trevena L. Decision aids for people facing health treatment or screening decisions. Cochrane Database of Systematic Reviews 2017, Issue 4. Art. No.: CD001431. https://decisionaid.ohri.ca/ds_tools.html
3. <https://www.healthdecision.org/tool#/tool/mammo>

Shared Decision Making and Clinical Utility

What is shared decision making?

The definition taken from the for Patients page of the Center for Shared Decision Making at Dartmouth Hitchcock Medical Center states that shared decision-making is “the collaboration between patients and caregivers to come to an agreement about a healthcare decision.” Although there isn’t one clear definition, it is about respecting a patient’s autonomy to allow them to make well informed decisions about the medical care they receive. To paraphrase a mentor, “you are the expert in medicine, the patient is the expert on their life.”

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

This begs the question, does engaging in shared decision-making change patient outcomes?

From the patient perspective:

There are multiple benefits to engaging in shared decision making for patients. Two areas where shared decision making appears to consistently perform better over “standard care” is in reducing decisional conflict, the uncertainty one feels when making choice, and increased satisfaction with decision making process. Widely variable patient populations all appear to appreciate the chance to be included more in the decision-making process, from teenagers with chronic disease to geriatric patients with dementia.

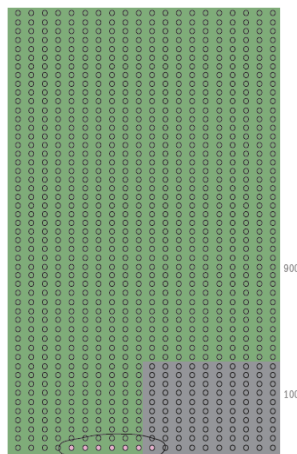
BENEFITS AND HARMS OF ASPIRIN OVER 10 YEARS

The primary benefit of aspirin is that it may help prevent a heart attack. The primary harm of aspirin is a risk of bleeding from the stomach that will require you to receive emergency care, receive blood transfusion, undergo endoscopy, and stay in the hospital for about 3 days, expecting a full recovery.

NO ASPIRIN

If 1000 people like you, DO NOT take aspirin...

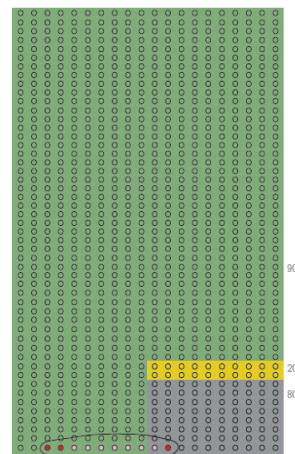
- 900 people DO NOT have a heart attack (green)
- 100 people DO have a heart attack (grey)
- 7 people DO experience bleeding that is NOT RELATED to aspirin (pink)



YES ASPIRIN

If 1000 people like you, DO take aspirin...

- 900 people DO NOT have a heart attack (green)
- 80 people DO have a heart attack (grey)
- 20 people AVOIDED a heart attack (yellow)
- 980 people experienced NO BENEFIT from taking aspirin
- 7 people DO experience bleeding that is NOT RELATED to aspirin (pink)
- 3 people DO experience bleeding RELATED to aspirin (red)



AVERAGE (10%) | © 2010 Mayo Foundation for Education and Research. All Rights Reserved.

Patient knowledge on their illness and available treatment options appears to trend toward improvement with shared decision making although there is some uncertainty if the improvement is clinically significant. It is unclear if shared decision making has any effect on quality of life due to the few studies available.

Overall, these factors combined are thought to contribute to a reduced likelihood of decision regret after a choice is made. It is important to note that many studies evaluating shared decision making exclude patients with low health literacy or significant barriers to follow-up care making it difficult to know if shared decision-making benefits these patients.

From the provider perspective:

It becomes less clear what benefits shared decision making provides from the provider perspective. Provider adherence to shared decision making has several barriers including the potential increase in time spent with patients necessary to engage in shared decision making with patients and the lack of knowledge of decision aids that promote the process. Additionally, some specialty specific concerns can arise, particularly with elective surgeries where the patient is presumed to have failed more conservative therapies prior to consulting with a surgeon.

Similarly, to patients, providers report greater satisfaction when engaging in shared decision making with patients. For outpatient encounters, shared decision making does not appear to increase overall visit length although generalizability to other clinical contexts is less clear. It is also unclear if the decisions patients make that deviate from standard care will have a beneficial, neutral, or negative effect on patient's health. One example of unclear benefits versus harm is engaging in shared decision making with the use of an aid in deciding whether to screen a woman for breast cancer. Despite the recommendation by several organizations that women in their 40s discuss its utility for them, there are concerns that by engaging in shared decision making, there was a 77% increase in women deciding to delay screening.

Where does that leave shared decision making?

It is difficult to parse out the benefits and outcomes associated with shared decision making. This is in part due to the lack of a universally accepted definition of shared decision making. The definition can range from the definition above provided the Center for Shared Decision Making at DHMC to the simple addition of a decision aid into the decision-making process. This can make comparing any outcomes across studies difficult although there seems to be overall agreement that there is an increase satisfaction with the decision-making process for both patients and providers. Additionally, despite uncertainties over the long-term outcomes for patients, shared decision making allows patients to make decisions that are congruent with their values.

While further research and an agreed upon definition of shared decision making is required to further understand its role in patient care, shared decision making is a valuable tool in an era of increasingly patient-centered care and in upholding patient autonomy.

References:

1. Baik, D., Cho, H. & Creber, R. M. M. Examining Interventions Designed to Support Shared Decision Making and Subsequent Patient Outcomes in Palliative Care: A Systematic Review of the Literature. *American Journal of Hospice and Palliative Medicine*® **36**, 76–88 (2018).
2. Boss, E. F. *et al.* Shared Decision Making and Choice for Elective Surgical Care. *Otolaryngology–Head and Neck Surgery* **154**, 405–420 (2015).

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

3. Burch, J. & Magalhães, P. V. How do decision aids affect the understanding and decisions of people facing health treatment or screening decisions? *Cochrane Clinical Answers* (2017). doi:10.1002/cca.1693
4. Center for Shared Decision Making. *Dartmouth* Available at: https://med.dartmouth-hitchcock.org/cshared_decision_making_toolkits.html. (Accessed: 24th January 2020)
5. Coronado-Vázquez, V., Gómez-Salgado, J., Monteros, J. C.-E. D. L. & García-Colinas, M. A. Shared Decision-Support Tools in Hospital Emergency Departments: A Systematic Review. *Journal of Emergency Nursing* **45**, 386–393 (2019).
6. Dobler, C. C. *et al.* Impact of decision aids used during clinical encounters on clinician outcomes and consultation length: a systematic review. *BMJ Quality & Safety* **28**, 499–510 (2018).
7. Ivlev, I., Hickman, E. N., Mcdonagh, M. S. & Eden, K. B. Use of patient decision aids increased younger women’s reluctance to begin screening mammography: a systematic review and meta-analysis. *Journal of General Internal Medicine* **32**, 803–812 (2017).
8. Kashaf, M. S., Mcgill, E. T. & Berger, Z. D. Shared decision-making and outcomes in type 2 diabetes: A systematic review and meta-analysis. *Patient Education and Counseling* **100**, 2159–2171 (2017).
9. Kew, K. M., Malik, P., Aniruddhan, K. & Normansell, R. Shared decision-making for people with asthma. *Cochrane Database of Systematic Reviews* (2017). doi:10.1002/14651858.cd012330.pub2
10. Martínez-González, N. A. *et al.* Shared decision making for men facing prostate cancer treatment: a systematic review of randomized controlled trials. *Patient Preference and Adherence* **Volume 13**, 1153–1174 (2019).
11. Peterson, E. B. *et al.* Impact of provider-patient communication on cancer screening adherence: A systematic review. *Preventive Medicine* **93**, 96–105 (2016).
12. Poprzeczny, A. J., Stocking, K., Showell, M. & Duffy, J. M. N. Patient Decision Aids to Facilitate Shared Decision Making in Obstetrics and Gynecology. *Obstetrics & Gynecology* **1** (2020). doi:10.1097/aog.0000000000003664
13. Riikonen, J. *et al.* Decision aids for prostate cancer screening choice: A systematic review and meta-analysis. *European Urology Supplements* **17**, (2018).
14. Scalia, P. *et al.* The impact and utility of encounter patient decision aids: Systematic review, meta-analysis and narrative synthesis. *Patient Education and Counseling* **102**, 817–841 (2019).
15. Wyatt, K. D. *et al.* Shared Decision Making in Pediatrics: A Systematic Review and Meta-analysis. *Academic Pediatrics* **15**, 573–583 (2015).

V.19 Health Literacy and Numeracy: Tactics to Improve Communication and Patient Understanding of EBM (Lily Greene, GSM4)

What is health literacy and numeracy?

There are many definitions of health literacy, which describe skills necessary to function in a health care environment. Healthy People 2020 defines it as the capacity for individuals to obtain, process, and understand health information required to make healthcare decisions.¹ A narrower definition, functional health literacy, refers to reading, writing, and numeracy skills required to make everyday health decisions. Health numeracy is a related concept that more narrowly focuses on the ability to assess, interpret, and act on numerical, graphical, and statistical information to make health decisions.² When we as healthcare providers conceptualize EBM, we must constantly consider not only how the newest evidence can be applied to our patients but also how we can effectively communicate the latest data on disease diagnosis and treatment in order to foster a rich dialog of shared-decision making. However, an individual patient's familiarity with health literacy and numeracy skills may impact shared decision-making and informed consent. We cannot assume that all patients have adequate literacy skills to directly engage with EBM as we are taught as healthcare providers. An extensive national assessment of health literacy in 2003 found that roughly 1/3 of US adults had low health literacy, and a more recent national survey found that 19% and 29% of US adults had low literacy and numeracy performances, respectfully.^{3,4}

How does health literacy affect health care engagement?

Low health literacy can affect patient engagement in the medical system. Patients with low health literacy often experience poorer health outcomes from less use of preventive care or difficulty interpreting health messaging.⁵ In the realm of treatment decision-making, a lack of complete understanding of the harms and benefits of treatment may lead to selecting a less suitable option or regretting the decision.⁶ Additionally, our study of EBM as students and providers makes us acutely aware of the need to recruit more diverse and representative populations in medical research. Increased diversity augments generalizability and a study's quality when applying the results to equally diverse and potentially vulnerable patient populations. However, previous studies have shown that patients with low health literacy have greater difficulty understanding the informed consent process and a decreased interest in participating in clinical research.⁷ Given this impact, there is a significant need to improve provider communication and patient understanding of EBM.

Tactics to improve patient-provider communication with EBM

To improve and tailor communication strategies to patients with low health literacy, the provider must first identify those who might benefit. This may seem obvious; however, it is worth noting that most providers overestimate a patient's level of health literacy, and many patients may downplay their level of knowledge due to associated stigma. So, obtaining an accurate assessment of a patient's comfort level with medicine and EBM topics may not be straightforward. There is currently no consensus on

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

the best method to assess health literacy. However, screening tools exist, like the S-TOFHLA (Short version of Test of Functional Health Literacy in Adults), which assesses reading and numeracy skills (Figure 1).⁸ Despite the ease of administration, screening tools alone are not always the perfect option, as significant stigma around literacy persists, and patients struggling with health literacy may feel discouraged from participating. One way to address this is using a short screening tool framed as a method of personalizing how to best deliver health information to the patient instead of detecting discomfort with literacy. Ideally, if a patient is identified as having lower health literacy, this could be signaled in the electronic health record in a non-stigmatizing way to help other providers identify which patients could benefit from tailored communication strategies.

Once identified, one strategy for addressing low health literacy is using patient-decision aids, which help patients understand a particular disease and provide relevant information on treatments. In utilizing patient decision aids, especially among patients with lower health literacy, it is vital not to provide overwhelmingly large amounts of information that may have opposing effects. Further information describing the implementation and creation of patient decision aids can be explored in the EBM Guide chapter titled “Decision Aids and Shared Decision Making” written by George S. Wang and Jesus Mendez Jr. Furthermore, when using electronic decision aids and information, it is essential to be aware of the patient's digital literacy skills. The use of technology may exclude patients with either limited access to or limited skill in accessing electronic information on the internet. Tailoring the medium and providing this information to the individual patient's skill set and comfort can help avoid this.

In addition to providing patients with appropriate information, altering communication strategies when framing EBM to patients can help increase patient understanding. EBM information can be framed as either a gain or a loss, where gain framing emphasizes the benefits of a particular decision and loss focuses on the harms involved.⁸ This is important to consider when presenting information to patients, as a risky treatment communicated with gain framing focused on increased survival may be more likely to be chosen than when the harms are emphasized, also called framing bias. In this way, our methods of communication can inadvertently steer patients in a specific direction. Therefore, best practice includes presenting gain and loss framing of the evidence simultaneously to prevent misinterpretation or undue influence. In addition, when presenting the results of a new study to a patient, it is important to remember that the type of result presented can send different messages to the patient about the overall impact of the intervention. Consider the following as an example of this point:

A 67-year-old male patient with a past medical history of obesity and coronary artery disease comes into your office to discuss starting a new medication he heard about on the news. He saw a story about new research showing the GLP-1 agonist semaglutide decreases the risk of adverse cardiovascular outcomes and wants to know if he should start it given his history of heart disease. You review the paper and plan how to discuss the results with the patient.

The study shows that patients who took semaglutide had a relative risk reduction of ~19% in the cardiovascular composite endpoint compared to placebo. This number was widely

written about in the news, and if presented with this number alone, your patient would likely think that the drug has a significant positive effect. However, since the RRR is a proportion, it can artificially inflate what could be an otherwise slight absolute difference in risk, which is the case in this study. Suppose instead, you presented that the semaglutide group had an ARR 1.5% in composite cardiovascular outcomes compared to placebo. In this case, the patient may need more convincing of the overall effect of the drug. Another way of explaining the results would be to use the number needed to treat, which in this case is 67. Explaining that 67 people would need to take semaglutide for one person see a positive benefit provides a more easily digestible explanation that the magnitude of the results may not be as significant as advertised in the news. Through this discussion, the patient and provider can determine he is not interested in starting semaglutide at this time. A further nuance in the discussion might relate to the use of composite outcomes, where the specific outcome of interest might not be achieved at all.

Tactics to improve patient understanding of EBM topics:

The EBM skills we develop as health care providers help us critically discern medical literature and how it informs decision-making. Though we may be able to offer patients recommendations based on the latest evidence, many patients lack familiarity and education with EBM skills, which can impair the effectiveness of shared decision-making. In addition, patients often seek medical information outside of their provider through the internet, and having the skill set to critically appraise health information is extraordinarily useful to allow for improved understanding and decision-making. One intervention to improve patient familiarity and understanding of EBM is through group and online courses focused on this topic. Studies have demonstrated the effectiveness of group courses teaching EBM, with participants reporting an increased comfort with handling health information and confidence in making correct health decisions.⁹ Despite their success, in-person multi-day courses can be more challenging to access and require considerable time investment for the patient. Web-based courses, which can be completed asynchronously, are likely a better solution. The US Cochrane Center currently provides a free web-based lecture-style course, "Understanding Evidence-Based Health Care: A Foundation for Action." In a survey study, participants who completed this program endorsed increased confidence in EBM topics like systematic reviews and how to search PubMed; however, it is worth noting that most participants in this study had an educational attainment of a bachelor's degree or higher.¹⁰ It is therefore unclear how well this educational training may be used in those with lower educational backgrounds.


With this limitation in mind, a research group in Japan designed a more accessible web-based EBM tool designed for laypersons. Instead of lecture-style instruction, their e-learning material consisted of scenarios where cartoon characters learn the fundamentals of EBM with true/false quizzes (Figure 2). The authors found that the learning content was of interest to non-health professions and, with a more game-based format, was enjoyable to complete.¹¹ In summary, these provide support for the idea of web-based learning tools for patients to increase familiarity with EBM techniques.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine


<p style="text-align: center;">Short Test of Functional Literacy in Adults STOFHLA READING COMPREHENSION</p> <p>HAND PATIENT THE READING COMPREHENSION PASSAGES TO BE COMPLETED. FOLD BACK THE PAGE OPPOSITE THE TEXT SO THAT THE PATIENT SEES ONLY THE TEXT.</p> <p>PREFACE THE READING COMPREHENSION EXERCISE WITH:</p> <p>"Here are some other medical instructions that you or anybody might see around the hospital. These instructions are in sentences that have some of the words missing. Where a word is missing, a blank line is drawn, and 4 possible words that could go in the blank appear just below it. I want you to figure out which of those 4 words should go in the blank, which word makes the sentence make sense. When you think you know which one it is, circle the letter in front of that word, and go on to the next one. When you finish the page, turn the page and keep going until you finish all the pages."</p> <p>STOP AT THE END OF 7 MINUTES</p> <p>PASSAGE A: X-RAY PREPARATION</p> <p>PASSAGE B: MEDICAID RIGHTS AND RESPONSIBILITIES</p>	<p>PASSAGE A</p> <p>Your doctor has sent you to have a _____ X-ray.</p> <p>a. stomach b. diabetes c. stitches d. germs</p> <p>You must have an _____ stomach when you come for _____.</p> <p>a. asthma b. empty c. incest d. anemia</p> <p>a. is. b. am. c. if. d. it.</p> <p>The X-ray will _____ from 1 to 3 _____ to do.</p> <p>a. take b. view c. talk d. look</p> <p>a. beds b. brains c. hours d. diets</p>
--	---

Figure 1: S-TOFHLA health literacy questionnaire example


Characters of the story




Mother "Haruko"
She is in her 70s and forgetful these days.



Wife "Yoshiko"
She is in her 40s and loves TV shopping.



"Mr. Natto-kun"
He is an office worker in his 40s and a member of the local sandlot baseball team. He is getting fat.




Daughter "Eri"
She is a high school student. She is interested in beauty and weight loss.

2. Evaluation of information on the internet 2 / 18 quizzes

Mr. Natto-kun asked, "Are there any good ways to effectively lose weight?"

His daughter Eri began checking it on her smartphone and showed him the information at the top of the search results.



Back Next


Quiz 2 / 18 quizzes

Mr. Natto-kun said, "It is not always true that the information at the top of the search results is correct."

True or false?

○

✗



Back


3 Answer 2 / 18 quizzes

Internet search is very convenient.

However, it is not always true that the information at the top of the search results is correct.

Besides, the information on the Internet may lack scientific evidence.

We cannot always believe everything on the Internet.



Back Learn more! Next

Figure 2: Example of graphics and questions used in EBM e-learning tool (adapted from Okabayashi S. et.al. E-Learning Material of Evidence-Based Medicine for Laypersons. *Health Lit Res Pract.* 2022

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

References:

1. Ratzan SC, Parker RM. Introduction. In: Selden CR, Zorn M, Ratzan SC, Parker RM, editors. In National Library of Medicine current bibliographies in medicine: Health literacy. Bethesda, MD: National Institutes of Health; 2000.
2. Golbeck AL, Ahlers-Schmidt CR, Paschal AM, Dismuke SE. A definition and operational framework for health numeracy. *Am J Prev Med.* 2005;29(4):375-376. doi:10.1016/j.amepre.2005.06.012
3. Cutilli CC, Bennett IM. Understanding the health literacy of America: results of the National Assessment of Adult Literacy. *Orthop Nurs.* 2009 Jan-Feb;28(1):27-32; quiz 33-4. doi: 10.1097/01.NOR.0000345852.22122.d6.
4. *Highlights of the 2017 U.S. PIAAC Results Web Report* (NCES 2020-777). U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Available at https://nces.ed.gov/surveys/piaac/national_results.asp
5. Nutbeam D, Lloyd JE. Understanding and Responding to Health Literacy as a Social Determinant of Health. *Annu Rev Public Health.* 2021;42:159-173. doi:[10.1146/annurev-publhealth-090419-102529](https://doi.org/10.1146/annurev-publhealth-090419-102529)
6. Hasannejadasl H, Roumen C, Smit Y, Dekker A, Fijten R. Health Literacy and eHealth: Challenges and Strategies. *JCO Clin Cancer Inform.* 2022;6:e2200005. doi:[10.1200/CCI.22.00005](https://doi.org/10.1200/CCI.22.00005)
7. Kripalani S, Heerman WJ, Patel NJ, et al. Association of Health Literacy and Numeracy with Interest in Research Participation. *J Gen Intern Med.* 2019;34(4):544-551. doi:[10.1007/s11606-018-4766-2](https://doi.org/10.1007/s11606-018-4766-2)
8. Hasannejadasl H, Roumen C, Smit Y, Dekker A, Fijten R. Health Literacy and eHealth: Challenges and Strategies. *JCO Clin Cancer Inform.* 2022;6:e2200005. doi:[10.1200/CCI.22.00005](https://doi.org/10.1200/CCI.22.00005)
9. Berger B, Gerlach A, Groth S, Sladek U, Ebner K, Muhlhauser I, Steckelberg A. Competence training in evidence-based medicine for patients, patient counsellors, consumer representatives and health care professionals in Austria: a feasibility study. *Zeitschrift Fur Evidenz, Fortbildung Und Qualitat Im Gesundheitswesen.* 2013;107(1):44-52.
10. Han G, Mayer M, Canner J, et al. Development, implementation and evaluation of an online course on evidence-based healthcare for consumers. *BMC Health Serv Res.* 2020;20(1):928. doi:[10.1186/s12913-020-05759-5](https://doi.org/10.1186/s12913-020-05759-5)
11. Okabayashi S, Kitazawa K, Kawamura T, Nakayama T. E-Learning Material of Evidence-Based Medicine for Laypersons. *Health Lit Res Pract.* 2022;6(4):e290-e299. doi:[10.3928/24748307-20221113-01](https://doi.org/10.3928/24748307-20221113-01)

Submitted 12/1/2023

V.20 Why is it so difficult to prove mortality as an endpoint? The challenge of studying mortality in the critical care setting (Ashley Baronner)

We often remark while rounding on the wards and in intensive care settings that there is no mortality benefit for some of our standard or guideline-driven interventions and treatments, and therefore conclude that they may be a waste of time. It has been suggested that as few as 15% of our medical interventions have been validated in the literature. However, we rarely stop to consider what it actually means when there is no proven mortality benefit. This especially true in critical care patients, for several reasons described below.

1. Mortality rates are decreasing over time due to ongoing improvements in medical care. This means that a larger sample size is required to detect changes in mortality because they are getting smaller overtime. This could be the difference between powering a study to detect a 10% absolute mortality difference versus a 2% absolute mortality difference, which would require a significantly increased power. These power calculations can be challenging to calculate when designing a study, as they are usually based on the prior mortality rates that have since declined.
2. Most studies are unlikely to show any change in mortality. Depending on patients' medical conditions and unique co-morbidities, they may be very likely to die despite any interventions made by their medical team versus very unlikely to die as long as they receive the standard of care. There are much smaller numbers of patients who fall into an intermediate category in which a specific intervention has the ability to dramatically change their mortality.
3. Interventions under study are often delivered too late to affect mortality. This is especially true in critical care patients, as the process of recruitment, enrollment and consent, randomization, and intervention takes time and it becomes too late in the disease process. In critical care patients, the first 24 hours of admission can be the most crucial for determining mortality, and RCT enrollment simply takes too long.
4. Clinical trials with a mortality endpoint often utilize a p value of <0.05 . This leads to trials that may show a significant outcome but are not reproducible or robust. These trials may be representing chance alone, particularly when studying heterogeneous intensive care patients. Additionally, if a trial is negative as a whole, or, as in a meta-analysis combines heterogeneous studies, it does not necessarily rule out a small mortality benefit, which can be meaningful to individual patients.
5. "Delta inflation" occurs when clinical trial investigators predict an unrealistically large improvement in mortality, leading to their studies being too small and insufficiently powered. This is more common in a field like critical care with large heterogeneity and rare case presentations that may not have clear diagnoses. In comparison, cardiology trials are often more defined, very large, and adequately powered for mortality benefits.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

For these and other reasons, studying mortality clearly has challenges that are especially prominent in critically ill patients. While it may be easier to select proximal endpoints such as ventilator free days, ICU free days, or discharge home, these are typically less important to our patients, and often unconvincing to clinicians. Cautious interpretation of secondary endpoints can allow us to continue to study mortality but derive other important data from these RCTs. For example, the ARDSNetwork published two studies indicating no significant effect of steroids or restrictive fluid strategy on overall mortality rates, but a positive effect on event-free days. Thus, it is important to continue to study mortality, and to not immediately reject RCT results that are unable to definitively prove a mortality benefit. This requires a close reading of the literature and a nuanced understanding of outcomes of meaning.

References:

1. Ospina-Tascón GA, Büchele GL, Vincent JL. Multicenter, randomized, controlled trials evaluating mortality in intensive care: doomed to fail?. *Crit Care Med*. 2008;36(4):1311-1322. doi:10.1097/CCM.0b013e318168ea3e
2. Harhay MO, Wagner J, Ratcliffe SJ, et al. Outcomes and statistical power in adult critical care randomized trials. *Am J Respir Crit Care Med*. 2014;189(12):1469-1478. doi:10.1164/rccm.201401-0056CP
3. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA*. 2002;288(3):358-362. doi:10.1001/jama.288.3.358
4. Johnson, Valen E. "Revised standards for statistical evidence." *Proceedings of the National Academy of Sciences* 110.48
5. Petros, A. J., J. C. Marshall, and H. K. F. Van Saene. "Should morbidity replace mortality as an endpoint for clinical trials in intensive care?." *The Lancet* 345.8946 (1995): 369-371.

V.21 How to assess treatment efficacy in solid tumor - an introduction to RECIST criteria – (Yuanzhen Cao)

Background

“In recent decades, the practice of cancer medicine and the technology of experimental cancer therapy have reached progressively higher levels of scientific sophistication.... There is, however, one essential element of this experimentation which is perhaps too frequently forgotten amidst such technical sophistry, namely, the actual measurement of the study end point. The culmination of most experimental therapeutic trials for solid tumors occurs when a man places a ruler or caliper over a lump and attempts to estimate its size. With this is introduced the inevitable factor of human error. Although the ultimate aim of therapy is increased survival, only few of our current approaches achieve that goal. To search for antitumor activity in a new modality by using survival as an endpoint is a far too complex and time-consuming effort and one that is frequently confused by the multiple therapies that may be attempted in any single patient.”

---Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer* 1976.

RECIST (Response Evaluation Criteria in Solid Tumors) is a set of standardized guidelines that provides a simple, consistent and pragmatic methodology to evaluate the activity and efficacy of new cancer therapeutics. In the 1960s, as research on cancer therapies and regimens started to proliferate, it became evident that different investigators might interpret treatment response (e.g. benefit) differently. Moertel et al. was the first to discuss this idea on *Cancer* in 1976, which marked the beginning of the modern drug assessment era. In 1981, the first standardized criteria for response assessment were published by the World Health Organization (WHO), which became the prototype for RECIST. Between the mid-1990s to 2000, an international collaboration which included the European Organization for Research and Treatment of Cancer (EORTC), National Cancer Institute of the United States, and the National Cancer Institute of Canada Clinical Trials Group, worked on further simplifying the response criteria. As a result, the RECIST criteria were first published in 2000 (Version 1.0). These criteria were subsequently updated in 2009 ([Verson 1.1](#)) and have become the gold standard for assessing the effectiveness of treatments in clinical trials. The criteria focus on solid tumors, such as those found in lung, breast, colon, and other organs.

Target lesions

RECIST allows for the identification of up to five target lesions per patient and two per organ, typically the largest measurable ones, to assess tumor burden quantitatively and track treatment response. Non-target lesions are evaluated qualitatively. For a lesion to be measurable, it must have a longest

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

diameter of at least 10 mm (or 15 mm for lymph nodes) on imaging studies such as CT scans or MRI. Prior to treatment initiation, baseline tumor measurements are obtained to establish the initial size and extent of the tumor.

Response Assessment

The main outcome of the RECIST criteria is the categorization of tumor response into four classes:

- Complete Response (CR): Disappearance of all target lesions.
- Partial Response (PR): At least a 30% decrease in the sum of the longest diameters of target lesions.
- Stable Disease (SD): Neither sufficient shrinkage to qualify as PR nor sufficient increase to qualify as Progressive Disease (PD).
- Progressive Disease (PD): At least a 20% increase in the sum of the longest diameters of target lesions, or the appearance of new lesions.

In regards to non-target lesions, they are generally not considered to contribute to the overall response assessment, but can provide valuable clinical information in determining treatment plan and prognosis. The key points are summarized in Table 1.

Other key terminologies used in response assessment

- Duration of response: measures the time from the first documented response (CR or PR) to disease progression or relapse is measured to assess how long the response to treatment lasts.
- Objective response rate: The percentage of patients whose RECIST results are classified as complete response or partial response.
- Progression free survival (PFS)

Application and limitation

RECIST criteria are closely related to treatment outcomes, and their reproducibility is generally acceptable when appropriate measures are taken in clinical assessments. Even though it might appear intuitive, evidence is supported by several large studies which demonstrated that a decrease in tumor size, as measured by RECIST criteria, is associated with improved overall survival (OS) and progression free survival (PFS). Conversely, an increase in tumor size is linked to worse outcomes. The precision of RECIST and of response categories has been studied extensively. Important factors associated with RECIST measurement reproducibility are the choice and number of target lesions which constitutes categorization of treatment response. Studies have shown that RECIST measurements have a reproducibility of about +/- 20% in multi-observer studies and +/- 10% in single observer studies. The expertise of the reader, how the lesions are defined, and their size also influence reproducibility,

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

emphasizing the significance of adhering to RECIST guidelines when choosing target lesions. To reduce variability, it is suggested that clinical trials include a central review with two readers and one adjudicator.

It is important to note that while the RECIST criteria have been widely used, they may not capture all aspects of tumor response. The presence of new lesions and progression of non-target lesions were found to be strongly associated with worse OS, emphasizing the importance of considering non-target lesions for assessing disease progression especially in clinical settings. Additionally, the initial version of RECIST did not consider bone metastases measurable due to the limitations of existing techniques in detecting bone marrow infiltration. With the updated RECIST 1.1, bone metastases with soft tissue masses ≥ 10 mm are now recognized as measurable target lesions. However, bone lesions without soft tissue involvement still remain unmeasurable according to RECIST criteria. As RECIST is not organ-specific, it might not capture the key parameters that are associated with survival outcomes in certain cancer types such as GIST and mCRC, in which case liver involvement is associated with poorer outcomes.

Applying RECIST criteria to evaluate treatment response is also debatable in certain focal treatments such as radiofrequency ablation, microwave ablation or cryoablation because they often leave a larger defect than the original lesion so that the lesions are considered unmeasurable. Another important concept is “pseudoprogression”, which describes the phenomenon of temporary increase in tumor burden observed in patients on immunotherapies. This is due to the fact that immune response triggered might initially cause inflammation and tumor swelling, thus delaying visible tumor shrinkage. In some cases, tumor bulk might not respond homogeneously and a mixed response can be observed, which could be challenging to categorize and can affect the objective response assessment.

Table 1. Summary of response assessment in RECIST 1.1

Overall Response	Target Lesions	Non Target Lesions	New Lesions
Definition	<ul style="list-style-type: none"> Lesions with longest diameter ≥ 10 mm and limits that are sufficiently well defined for their measurement to be considered reliable Lymph nodes: measurement of short axis, target lesion if short-axis measures ≥ 15 mm Maximum number of selected target lesions 5/patient and 2/organ 	<ul style="list-style-type: none"> Lesions that are too small (< 10 mm) Lesions for which measurement is considered unreliable as their limits are difficult to define (bone or leptomeningeal lesions, ascites, pleural or pericardial effusion, lymphangitic carcinomatosis etc.) Measurable lesions not selected as target lesions Lymph nodes: measurement of short axis, non-target lesion if $10 \text{ mm} \leq \text{short-axis diameter} < 15 \text{ mm}$ Levels of tumour markers $>$ normal (if relevant and predefined) 	
Complete response (CR)	<ul style="list-style-type: none"> Disappearance of all target lesions and all nodes have short axis < 10 mm 	<ul style="list-style-type: none"> Disappearance of all non-target lesions and normalisation of tumour marker levels No progression 	<ul style="list-style-type: none"> No
Partial response (PR)	<ul style="list-style-type: none"> $\geq 30\%$ decrease in the sum of target lesions taking as reference the baseline sum 		<ul style="list-style-type: none"> No
Stable disease (SD)	<ul style="list-style-type: none"> Neither response nor progression 	<ul style="list-style-type: none"> Persistence of one or more non-target lesions and/or tumour marker levels $>$ normal 	<ul style="list-style-type: none"> No
Progressive disease (PD): response is PD if at least one category of lesions meets progression criteria	<ul style="list-style-type: none"> $\geq 20\%$ increase in the sum of target lesions taking as reference the smallest sum measured during follow-up (nadir) and ≥ 5 mm in absolute value 	<ul style="list-style-type: none"> 'Unequivocal' progression (assessed qualitatively) in lesion size (an increase in size of a single lesion is not sufficient) 	<ul style="list-style-type: none"> Yes [appearance of new unequivocally metastatic lesion(s)]

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Conclusion

RECIST criteria were developed for clinical trials to ensure that patients are classified in a comparable manner, taking into account the variability in tumor measurements. These criteria are widely used in clinical trials and accepted by regulatory agencies as a standardized tool for evaluating treatment effectiveness. Although there are some limitations, the scientific basis for using RECIST-based surrogate endpoints to approve anticancer drugs remains valid. The reproducibility of RECIST is influenced by factors such as reader experience, target lesion selection, and the detection of new lesions. Adequate training of radiologists is crucial for improving its application. Overall, RECIST serves as a common language that is extremely useful in clinical research between oncologists and imaging experts when there is a full understanding of how measurements are made, what they represent, and their inherent limitations.

References

- Fojo, Antonio T., and Anne Noonan. 2012. "Why RECIST Works and Why It Should Stay—Counterpoint." *Cancer Research* 72 (20): 5151–57.
- Fournier, Laure, Lioe-Fee de Geus-Oei, Daniele Regge, Daniela-Elena Oprea-Lager, Melvin D'Anastasi, Luc Bidaut, Tobias Bäuerle, et al. 2021. "Twenty Years On: RECIST as a Biomarker of Response in Solid Tumours an EORTC Imaging Group - ESOI Joint Paper." *Frontiers in Oncology* 11: 800547.
- Jaffe, C. Carl. 2006. "Measures of Response: RECIST, WHO, and New Alternatives." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 24 (20): 3245–51.
- Nishino, Mizuki, Jyothi P. Jagannathan, Nikhil H. Ramaiya, and Annick D. Van den Abbeele. 2010. "Revised RECIST Guideline Version 1.1: What Oncologists Want to Know and What Radiologists Need to Know." *American Journal of Roentgenology* 195 (2): 281–89.
- Moertel, C. G., and J. A. Hanley. 1976. "The Effect of Measuring Error on the Results of Therapeutic Trials in Advanced Cancer." *Cancer* 38 (1): 388–94.

Submitted 7-26-2023

V.22 Race and Ethnicity in EBM and Biomedical Research (Maya DeGroot, GSM4)

Introduction and Executive Summary:

Racism in biomedical research has contributed significantly to race-based medicine, “the system by which research characterizing race as an essential biological variable, translates into clinical practice, leading to inequitable care.” Race, a social and power construct, has been used as a proxy for genetic variation despite the demonstrated genetic heterogeneity within racial groups. As part of practicing Evidence-Based Medicine (for Life!), we must learn to critically assess how race and ethnicity are categorized and utilized in medical research and how this informs our practice. Racism must not be ignored, and studies should contribute to **race-conscious medicine** which “emphasizes racism, rather than race, as a key determinant of illness and health” in order to mitigate health inequities. The mechanism in which research contributes to racial health inequities is summarized in the figure below from an article published in the Lancet in October, 2020.¹

The categorization of race and ethnicity in research is very inconsistent and requires improved transparency and standardization. In 2003, Kaplan and Bennett published a guide in JAMA for researchers on the use of race and ethnicity in research which recommends that researchers include at least the following in their studies:

1. Specify the reason for using race or ethnicity when that data is presented
2. Racial and ethnic categories should be described, and the collection method justified
3. All relevant variables including social class should be included in the analysis²

The guide below is adapted from their suggested guidelines. It is meant as a guide for how to assess the ways in which research studies use or misuse categories of race and ethnicity and how this informs our interpretation and practice.

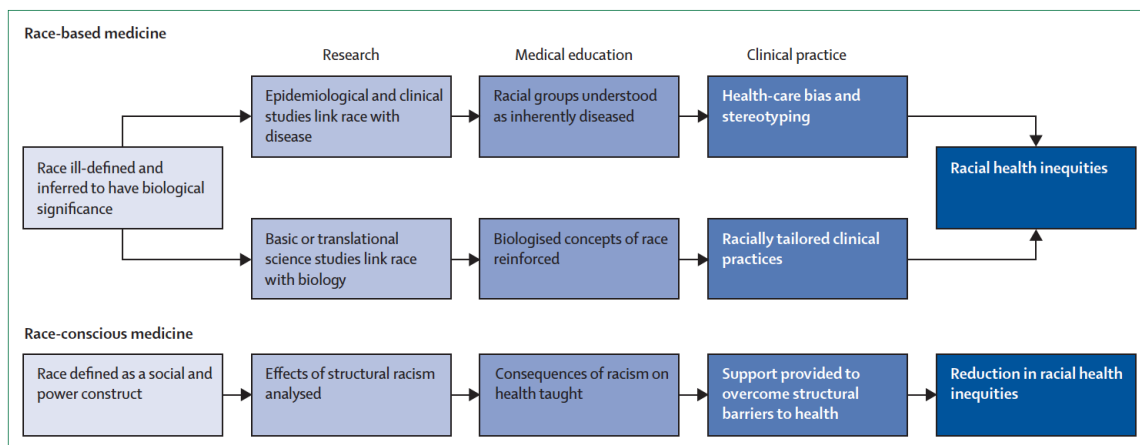


Figure: How race-based medicine leads to racial health inequities

An alternative approach to race-conscious medicine; defined as medical practice and pedagogy that accounts for how structural racism determines illness and health.

A Brief Note on the History of Racism in Biomedical Experimentation

Biomedical research and medical institutions have a long history of exploitation of and experimentation on Black people in the United States. Harriet Washington chronicles examples of these widespread atrocities in her 2006 book, *Medical Apartheid: The Dark History of Medical Experimentation on Black Americans from Colonial Times to Present*. The injustices and abuse have been widespread, including experimentations on slaves, birth control testing and development targeting Black communities as part of the eugenics movement, radiation experiments and even biological weapon development.³ Fortunately, there have been significant improvements in the protection of research subjects, including Black subjects in recent years. However, the exploitation of Black communities continues overseas as pharmaceutical companies and researchers search for new populations for testing.

Washington describes a seeming contradiction in the current climate of race and medical research in the United States: she encourages Black people in the US to participate in research studies to address the huge racial health inequities and rightly cautions black people to be wary of research abuses since although they are rare now, “the potential for exploitation and abuse still looms.”³ The absence of Black patients from important therapeutic research is problematic and rooted in historical research realities.³ It is important that as medical students and physicians, we acknowledge the historical and current racism in our medical institutions and research. We must also develop a framework for assessing whether study investigators are contributing to harmful race-based medicine or valuable race-conscious medicine. This chapter is meant to serve as a starting point for an approach to assessing the use of race, a powerful social construct with devastating health implications, in biomedical research studies.

1. **If race or ethnicity is used as a study variable, do the investigators specify the reason(s)?**

Since May 2000, The Uniform Requirements for Manuscripts Submitted to Biomedical Journals has required an explanation for the use of race in biomedical studies.² When scientists use social categories like race, there is the risk of this being interpreted as validating these categorizations. Thus, this explanation is important to include due to the risk of implying or reinforcing ideas that health disparities are caused by race itself rather than racism and specific mechanisms of racism that lead to the health disparities.²

2. **How are race and ethnicity being categorized and is this method justified?**

There exists a lot of variability in research around race and ethnic group categorization and data collection. Authors should describe the way in which individuals were assigned to racial or ethnic categories. Are they reporting the subject’s self-identified race, perceived race (what others believe a person to be) or reflected race (the race a person believes others assume them to be)?⁴ If identification was self-reported then “authors should specify whether individuals answered an open-ended question or chose from a fixed set of categories.”

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

One major challenge that researchers face is codifying the social constructs of race and ethnicity. In 1977, the Office of Management and Budget (OMB) set standards that have been since modified in which federal agencies must ask individuals to select 1 or more races when self-identifying and the five minimum categories must be: American Indian/Alaska Native, Asian, Black/African American, Native Hawaiian/Other Pacific Islander, and white. OMB provides two options for ethnicity: “Hispanic/Latino” and “not Hispanic /Latino.”²

These categorizations have many limitations, including that these identities are not static and that researchers encounter statistical challenges when including biracial or multiracial identities³. Bonham et al. describes in a 2018 JAMA article that the 2016 National Human Genome Research Institute and National Institute on Minority Health (NHGRI) and Health Disparities (NIMHD) workshop extensively discussed the use of self-identified race and ethnicity data in genomics, biomedical and clinical research and implication of the use for minority health and health disparities.⁴ They call for researchers to “increase the scientific rigor in collecting such data, especially in clinical settings” and that they need to collect data that reflect the multidimensional aspects of identity with regard to race, ethnicity, socioeconomic status (SES) and geographic ancestry.⁴ Currently there is limited standardization that prevents comparing data across studies and scientists are being called to improve and standardize the way in which such data is collected.⁴

Moore published an article in JAMA Ophthalmology in 2020 in which they assessed the frequency and use of race and ethnicity data in the ophthalmology literature in 2019. He found that 88% of ophthalmology articles that year reported patient age and sex but only 43% of studies reported race and/or ethnicity. Of those articles, only 13% described in the methods or results how these categories were determined.⁵ There is clearly a dearth of race and ethnicity reporting in ophthalmology and medical research and even when this information is included, it is very inconsistent.

3. Are the investigators using race or ethnicity as a proxy for genetic variation and upholding biological theories of race?

Race should never be used as a proxy for genetic variation. Genetic diversity within socially defined racial groups is greater than variation between these groups. Observed differences between groups in studies should consider all relevant factors that could be contributing to the differences, including racism and SES (see question 5 below). Kaplan and Bennet emphasize that “statements about genetic differences should be supported by evidence from gene studies. Genetic hypotheses should be firmly grounded in existing evidence, clearly stated, and rigorously tested.”²

4. Do the investigators distinguish between race as a risk factor and race as a risk marker when they state the hypothesis and describe the study results?

Membership in a racial group may be a risk marker for a particular group but membership in the group is not the risk factor, and investigators need to explicitly state this distinction. They should explain that the cause of the health disparity is not race itself but rather factors such as racism, lack of access to quality health care, or other societal factors that disproportionately impact the health outcomes in this racial group as compared to another racial groups.³ It is important to critically assess how racial inequities are explained and described in studies. Are they ascribed to the racist policies and societies that cause and perpetuate these inequities?

5. Do interpretations of racial or ethnic differences consider all relevant factors including racism and discrimination, social class, SES, personal or family wealth, environmental exposures, insurance status, age, diet and nutrition, health beliefs and practices, educational level, language spoken, religion, tribal affiliation, country of birth, parents' country of birth, length of time in the country of residence, place of residence, and/or zip code?

The study should try to include all relevant variables such as educational attainment, income, geographic residence, and other factors listed above, in addition to race and ethnicity. Researchers may encounter limitations and constraints when gathering this data and such constraints should be acknowledged. Racial health disparities, such as higher maternal mortality rates in African American women as compared to white women in the U.S., exist regardless of SES or education level and this is due to racism. Researchers need to do appropriate adjustments for all these factors. Kaplan and Kaplan described that “because lack of adjustment for SES or social class is the most important potential source of bias in studies of racial/ethnic differences, researchers should make every effort to adjust for conceptually relevant measures of SES or social class when comparing racial/ethnic groups. Unadjusted findings should be clearly labeled as such, and in general they should be reported in conjunction with adjusted findings for comparison purposes.”² Researchers should include SES or class diversity within racial groups and since SES and class may not be comparable across population groups, at least two SES measures are recommended to account for this. Race or ethnicity should never be a proxy for SES or social class since this is rooted in and contribute to racist ideas and stereotypes.²

Bonham emphasizes the need for “consensus about use of race, ethnicity, [social determinants of health], and ancestry data in study design, interpretation of results, publications and medical care.”⁴ We need to expand beyond traditional categories to explain population differences in order to understand how social, demographic, and biological factors affect health.

6. Do the authors use terminology that is stigmatizing, reflect unscientific classification systems and/or imply that race or ethnicity is an inherent, immutable attribute of an individual?

The language that authors use to discuss race and ethnicity is critical. Tables and figures should have footnotes that explain how racial and ethnic categories were defined and how individuals were assigned to them. If tables and figures include terms such as “race,” “ethnicity” or “race/ethnicity” then there should be caveats and explanations for these terms and authors should use more precise terms like “self-reported race” or “race category³.” Investigators should always specify that these racial and ethnic categories are not fixed identities of individuals.

How broad or specific are the racial or ethnic groups described? More specific terms to describe groups are preferable such as “Southeast Asian” instead of “Asian” and “Mexican American” or “Cuban American” instead of “Hispanic or latinx.” The term “caucasian” should not be used since the term originated from racial classifications rooted in “scientific racism, the false idea that races are naturally occurring, biologically ranked subdivisions of the human species and that Caucasians are the superior race⁶.” If the term “Caucasian” is used in reference to a previous study that used the term, then it should be put in quotations. The term “non-white” is never acceptable for many reasons including that it implies that the white population is normative.²

7. Is the race or ethnicity of a subject population characterized? Which racial or ethnic groups has been included or excluded?

NIH-funded researchers are required to use OMB census categories to report race and ethnicity in any research conducted with NIH funding to demonstrate inclusion of a diverse study population⁷. Although this requirement increases reporting of racial and ethnic categories, it does not necessarily impact the racial or ethnic recruitment process and study population⁸. When assessing the quality of a study, the reader must consider who is included in and excluded from the study. Are Black people, for example, under or overrepresented in the study population? To answer this question, one typically requires context. For example, for this elective I evaluated the ACTT-1 study on remdesivir in hospitalized patients with COVID-19, and when I was assessing the patient population, I categorized the study as adequately including participants of different races. However, Goldman et al. points out in a letter to the NEJM that since COVID-19 is disproportionately infecting Black, American Indian/Alaska Native and latinx folks, this study is severely underrepresenting these patients in the therapeutic trial⁹. Harriet Washington makes it clear in *Medical Apartheid* that there is a huge need for Black representation in therapeutic trials prior to COVID-19. Washington describes that “African Americans desperately need the medical advantages and revelations that only ethical, essentially therapeutic research initiatives can give them” and “history and today’s deplorable African American health profile tell us clearly that black Americans need both more research and more vigilance.”³ When assessing the quality of a study, we as the astute clinicians and readers must assess the study population very carefully to ensure that racial and ethnic groups are appropriately represented.

Additional questions to consider:

8. Does this study utilize race-based measurements, estimations or calculations?

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

9. Is race or ethnicity being controlled for in the statistical analysis and why?
10. What is the racist history of research and experimentation in the subspecialty of medicine? Does this specialty, publication and/or investigators recognize and acknowledge this history?

As we work towards racial equity at all levels of our society, medicine must reckon with and recognize the significant role of this profession and institution in racism. All of us are expected to acknowledge and actively work towards undoing the adverse effects of racism on health and health outcomes in this country, as well as internationally. American history, the current COVID19 pandemic, chronic illness disparities, our educational and prison policies, law enforcement and economic practices; all are emerging to face our society, and ask the questions required to repair the damages.

***Note: I would appreciate any feedback on this chapter. Please contact me at degroote.maya@gmail.com with thoughts, suggestions, or concerns. I invite others to make revisions and additions as part of your contribution to the EBM guide.*

References:

1. Cerdeña JP, Plaisime MV, Tsai J. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. *Lancet*. 2020 Oct 10;396(10257):1125-1128. doi: [10.1016/S0140-6736\(20\)32076-6](https://doi.org/10.1016/S0140-6736(20)32076-6)
2. Kaplan JB, Bennett T. Use of race and ethnicity in biomedical publication. *JAMA*. 2003 May 28;289(20):2709-16. doi: [10.1001/jama.289.20.2709](https://doi.org/10.1001/jama.289.20.2709)
3. Washington, H. *Medical apartheid: the dark history of medical experimentation on black americans from colonial times to present*. New York: Doubleday; 2006
4. Bonham VL, Green ED, Pérez-Stable EJ. Examining How Race, Ethnicity, and Ancestry Data Are Used in Biomedical Research. *JAMA*. 2018 Oct 16;320(15):1533-1534. doi: <https://pubmed.ncbi.nlm.nih.gov/30264136/>
5. Moore DB. Reporting of Race and Ethnicity in the Ophthalmology Literature in 2019. *JAMA Ophthalmol*. 2020 Aug 1;138(8):903-906. doi: [10.1001/jamaophthalmol.2020.2107](https://doi.org/10.1001/jamaophthalmol.2020.2107).
6. Mukhopadhyay, CC. Pollock, M. *Everyday antiracism: getting real about race in school*. New York: The New Press; c2008. Chapter 3, Getting rid of the word "caucasian"; p. 12-16.
7. National Institutes of Health. [NIH policy and guidelines on the inclusion of women and minorities as subjects in clinical research](#). Accessed December 21, 2020.
8. Callier SL. The Use of Racial Categories in Precision Medicine Research. *Ethn Dis*. 2019 Dec 12;29(Suppl 3):651-658. doi: [10.18865/ed.29.S3.651](https://doi.org/10.18865/ed.29.S3.651)
9. Goldman JD, Osinusi A, Marty FM. Racial Disproportionality in Covid Clinical Trials. *N Engl J Med*. 2020 Dec 17;383(25):2486-2487. doi: [10.1056/NEJMc2029374](https://doi.org/10.1056/NEJMc2029374)

Submitted 1/8/21

V.23 The Translational Highway- narrowing the gap between research and practice (Shantum Misra)

It has been reported that it can take up to 17 years for just 14% of novel scientific discoveries to be implemented in clinical practice. Unfortunately, the transition from research bench to clinical bedside has several roadblocks and these have often been described as a “translational highway.” Several authors have explored this highway, but often the focus hones in on the path researchers take to transition material from the research bench to a clinical investigation. There is less focus on the latter part of the highway, where the speed limit is slower: transitioning from clinical investigations to clinical practice.

What are the impediments to practical adoption of literature and how feasible is it to practice evidenced based medicine (EBM) when the evidence is not yet recommended by clinical guidelines? To answer this question, it’s first important to explore the usability of data published from clinical trials within a local context.

Application of EBM in Local Context

When researchers embark on conducting a clinical trial, they aim to represent the population at large. Metrics are used to standardize demographic variables so as to most effectively depict the patients who may actually benefit from an intervention. Unfortunately, quite often the implementation of a certain intervention on paper is not easily translatable to real life. Some strategies may be more generalizable than others (i.e., data to support hand hygiene is widely applicable) however often the vast majority of healthcare is being implemented in local, community clinics, private offices, and with general practitioners. At this juncture, the highway divides into multiple exits that can lead to sub-specialists and more narrow care. The adoption of EBM to the local context is vital in evaluating the latter part of the translational highway; it is important to consider the differences and similarities between clinical trials and real life.

The Physician's Role

As mentioned above, there are some clinical strategies that are generalizable to the population at large without significant impediment. Yet review of history shows that even these highly generalizable practices are not sufficiently implemented. For instance, it is widely accepted that Aspirin has a cardioprotective effect as a secondary prevention tool for patients with a history of cardiovascular disease. Despite this, Stafford et al. found that aspirin was being prescribed for just over one third of patients with coronary disease despite no contraindications to its use. Similarly, it took nearly 18 years for angiotensin converting enzymes to be accepted into clinical practice for treating left ventricular systolic dysfunction. And even though the first studies on the efficacy of beta-blockers in congestive heart failure were published in 1989, once again the strategy was not widely implemented until 2011. These examples illustrate that even when clinical trials show efficacy of therapeutic agents, numerous factors play a role in hampering its deployment into clinical practice. Is it the physician's role to be aware of the almost continuous changes that are occurring in medical practice? On the pathway of a research idea to picking up a medication at a pharmacy, where are the roadblocks and what can be done to expedite the process? The physician, particularly front-line physicians (i.e., internists, primary care, and family medicine) are tasked with the job of analyzing, reviewing, adopting, and implementing strategies in several areas of medicine. It is a daunting task that requires even the most skilled providers to be mindful.

Therapeutic Inertia and Medication Adherence

Therapeutic inertia (TI) is often defined as a failure to add or increase therapy when treatment goals are unmet. Therapeutic inertia can be defined numerically, as per the original definition in a paper by Okonofua et al in which:

$$\frac{h}{v} - \frac{c}{v}$$

Where h is the number of visits with an uncontrolled condition, v is the total number of visits, and c is the number of visits in which a change was made. As h increases and c decreases (meaning patients have more visits with an untreated condition) then the TI increases resulting in worse outcomes. In his paper, Dr. Okonofua demonstrated the large burden of TI in hypertension and the significance of reducing TI to improve blood pressure control.

Overcoming therapeutic inertia has been fundamental in the field of endocrinology in order to improve outcomes for patients living with diabetes. As recently as 2019, the American Diabetes Association released a statement in which they reported that, despite all the technological and pharmaceutical advancements in treating diabetes mellitus, nearly half of patients still have inadequate glycemic control. TI plays a role in the failure to adopt clinical guidelines into practice. Although it is not the sole contributor, it not only affects the well-being of patients but also the way patients view their providers.

The degree to which patients are invested in their own care is related to the degree providers are invested in their patients' care. However, the roadblocks mentioned earlier include patient willingness to participate in their own care. Granted, patients may not be aware of the cardioprotective benefits of Aspirin in secondary prevention, but even if a provider were to prescribe Aspirin, the onus would be on the patient to take the medication. The World Health Organization has reported that medication adherence can often have more of a direct impact on patient health than the medication itself. Nonadherence to medications can result in treatment failure, mortality, and re-hospitalizations. All of these contribute directly and indirectly to increased medical cost.

Steps to Reduce Traffic

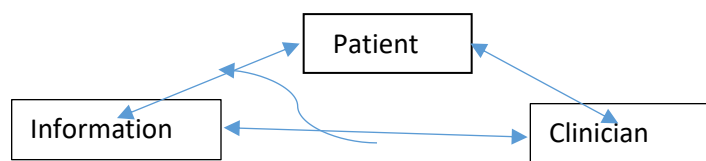
So, one can see how therapeutic inertia and medication non-adherence are both roadblocks to adopting clinical practice. So, what can be done to correct this? Communication is key. The barrier to improved healthcare outcomes is not necessarily the time it takes for a research idea to go through the rigorous process of becoming an accepted practice; more so, it is the adoption of that practice into society.

Communication is an essential tool in the physician's armamentarium that can help improve outcomes. Didactics for all providers, annual, mandatory seminars, and more education opportunities are necessary for providers to stay in tune with the incredibly fluid field that is medicine. Furthermore, the power of communication is important in improving medication adherence as well. Obesity is a common example in which studies have shown that the simple act of discussing weight loss in a meaningful way results in dramatic weight changes for patients. One study found that overweight and obese patients who were told by their physician that they need to modify their lifestyle to lose weight were more likely to report a 5% weight loss in one year.

The application of EBM is a multi-faceted process that requires the participation of multiple parties in order to improve patient care. Data shows that although there are several roadblocks in the highway that leads from the research bench to the clinic, a large portion of the burden is at the end of the road: with providers and their patients. Small steps, such as improved communication and education, can help to mitigate this burden.

A useful visual to illustrate the relationship between the clinician, the patient, and information captures the responsibilities and characteristics of each component required for communicating risks and benefits. Promoting and advocating adherence, deep knowledge of both the literature as well as the patient, and facilitating the patient's understanding of the evidence are core elements of this triad.

The patient brings individual knowledge, attitudes, and beliefs. The clinician brings knowledge, attitudes, fidelity to patient care, commitment to quality and the ability to translate information thereby facilitating knowledge transfer. The information must be accessible, valid, and helpful in a shared decision context.



Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Of course, this triad is not the only influence when considering how to bridge the gap between research and practice. But it's a good start. It is a step that we should be willing to take for the greater good of our patients and our own professional development.

References:

1. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *J R Soc Med.* 2011;104(12):510-520. doi:10.1258/jrsm.2011.110180
2. Schwartz K, Vilquin JT. Building the translational highway: toward new partnerships between academia and the private sector. *Nat Med* 2003;9:493-495
3. Stafford RS, Radley DC. The underutilization of cardiac medications of proven benefit, 1990 to 2002. *J Am Coll Cardiol* 2003;41:56-61
4. Waagstein F, Caidahl K, Wallentin I, Bergh CH, Hjalmarson A. Long-term beta-blockade in dilated cardiomyopathy. Effects of short- and long-term metoprolol treatment followed by withdrawal and readministration of metoprolol. *Circulation.* 1989; 80:551-563.
5. Okonofua EC, Simpson KN, Jesri A, Rehman SU, Durkalski VL, Egan BM. Therapeutic inertia is an impediment to achieving the Healthy People 2010 blood pressure control goals. *Hypertension.* 2006;47(3):345-351. doi:10.1161/01.HYP.0000200702.76436.4b
6. Khunti K, Davies MJ. Clinical inertia-Time to reappraise the terminology?. *Prim Care Diabetes.* 2017;11(2):105-106. doi:10.1016/j.pcd.2017.01.007
7. Brown MT, Bussell JK. Medication adherence: WHO cares? *Mayo Clin Proc.* 2011;86(4):304-314.
8. Pool AC, Kraschnewski JL, Cover LA, et al. The impact of physician weight discussion on weight loss in US adults. *Obes Res Clin Pract.* 2014;8(2):e131-e139. doi:10.1016/j.orcp.2013.03.003

Submitted 1/23/2021

V.24 A new therapy gains FDA approval, then what? (Jon Pirruccello)

Previous chapters of this book have described the pathway a new therapy takes through the FDA. From preclinical studies, to phase I safety, phase II efficacy, phase III comparability and phase IV post-marketing surveillance; the process is long, expensive, and arduous. The success and/or failure of a drug hinges on a highly critical statistical review throughout the process. After FDA approval, how does a new therapy become available to the insured patient?

After phase III, the following occurs:

To start, the vast majority of new medications are covered under health insurance plans –either government sponsored (Medicare/Medicaid) or private (Insurance companies like Harvard Pilgrim, Anthem, etc.). These plans, referred to as “payers” typically contract out their pharmacy plans to third party companies called Pharmacy Benefit Managers (PBMs). When a new drug gains approval, these PBMs carefully review the safety and efficacy data on the new therapy and decide whether or not they are going to cover the medication. They may decide to cover the therapy but add it to a certain “tier” on their formulary. In other words, they may require the consumer to pay more for the medication if they (the PBM) feel the evidence for the medication is not overwhelmingly strong, or if their projections indicate they are at financial risk. The PBM may also require a patient to try a certain therapy prior to paying for the new therapy. For example, a patient may be required to try simvastatin before the PBM pays for a newer statin. This process is called a “prior authorization (PA)”. The formulary committees of these PBM companies typically involve clinicians, pharmacists, nurses and financial analyst. Economic data is carefully intertwined with safety/efficacy data in order to determine whether the drug should be added to their formulary or not.

Two of the nation’s largest pharmacy benefit managers are Express Scripts and CVS Health (yes, the same CVS with drugstores on every corner). These PBMs contract directly with individual pharmacies to reimburse for drugs dispensed to individual patients. You may have heard some controversy surrounding PBMs negotiating rebates from manufacturers of the drugs. PBMs negotiate with manufacturers to determine drug price. These rebates are generally not disclosed publicly. Further, the rebates are not passed onto the insurance company. Some policy makers believe PBMs should be compelled to pass on the discounts to health insurers who could then pass on the savings to individuals in the form of lower premiums for their health insurance. This is beyond the scope of this chapter.

A large and evolving field is Health Economics and Outcomes Research (HEOR). Essentially, HEOR companies, or HEOR divisions of large pharma companies (manufacturers) consolidate the efficacy, safety, and cost of a drug/therapy into an appealing package and present it to governments, payers, health ministries, and hospital systems in order to ensure that decision makers are fully informed on their newly developed therapy. This is all done in an attempt to increase the utilization of their product. In a sense, HEOR companies help decision makers evaluate the economic, clinical and hard to measure costs or benefits of the new therapy.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

This is a cursory explanation of the post FDA approval path of a new drug. I hope it gives you a general understanding of the rigorous clinical and economic analysis a new therapy undergoes post approval. In short, the efficacy and safety data is not only analyzed by FDA regulators and future prescribers, but by the policy makers and future payers of the new medication. The transparency of this process to the eye of individual consumers is quite lacking, unfortunately, as the public has weak representation at the table.

[Deciding Which Drugs Get Onto the Formulary: A Value-Based Approach - ScienceDirect](#)

[How do insurers decide what medicines to pay for? - Business Insider](#)

[Pharmacy Benefit Managers and Their Role in Drug Spending | Commonwealth Fund](#)

[ISPOR - About HEOR](#)

Submitted 2/2021

V.25 Emergency Use Approval: What, How and Why it is used (Angela Lee, GSM4)

Chapter 61 of this guide discusses what exactly occurs when a new drug gains FDA approval. As noted, this process is arduous to ensure new treatments meet the utmost standards of scientific rigor. However, questions arose when, during the midst of the COVID pandemic, Moderna, Pfizer and Johnson & Johnson all received **Emergency Use Approval (EUA)** for their vaccines. This naturally raises queries into what exactly an EUA is and whether EUAs meant these vaccines were truly safe for patients. This chapter will provide a basic overview of an EUA and answer the question of whether treatments with EUAs are truly safe.

I. What is the definition of an EUA?

According to the FDA, an EUA is “a medical countermeasure used to combat chemical, biological, radiological, nuclear, and infectious disease threats, and are issued by the FDA during public health emergencies to facilitate access to drugs, diagnostic tests, or other essential medical products when there are no adequate, approved, and available options.”¹

II. OK...but what *exactly* does that mean?

In the simplest terms, an EUA is exactly what the term sounds like: it allows the medical community to use unapproved medical products in a public health emergency. Medical products include not just potential drugs but also unapproved diagnostic tests, medical devices, and vaccines. EUAs also allow currently existing medical products to be used for purposes that they were not originally intended for. An example of the latter is hydroxychloroquine; at the height of the COVID-19 pandemic, hydroxychloroquine was touted as a possible treatment for patients suffering from severe COVID-19.ⁱ Based on the limited evidence at the time, the FDA granted an EUA for the use of hydroxychloroquine in these patients, even though the only indications for its use prior to this had been for lupus and rheumatoid arthritis. (It should be noted, that this EUA was later rescinded).

III. Why was the EUA even created? What is its history?

To understand why the EUA was created, a brief history is instructive. In response to the 9/11 terrorist attacks in 2001, the federal government created and passed the Project BioShield Act in 2004. This act amended Section 564 of the pre-existing Federal Food, Drug and Cosmetic Act of 1938, allowing the FDA Commissioner to authorize unapproved medical products during a declared federal emergency. Thus, the original purpose of the EUA was designed in response to biological and/or chemical weapons used during possible terrorist attacks. However, while this was its original purpose, future pandemics were also considered appropriate for EUAs.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

While the EUA was approved in 2004, it was not until the 2009 H1N1 Flu that the FDA issued multiple EUAs in a single given year. The 2009 H1N1 Flu led to the writing and approval of The Pandemic and All-Hazards Preparedness Reauthorization Act of 2013 (PAHPRA). This act further amended the EUA, enabling the FDA to prepare for and prevent a public health emergency rather than use it purely in response to one. It also allowed EUAs to be issued in response to general threats that endangered either the American public's health and/or posed a significant threat to national security. Prior to this, the EUA could only be issued in response to a specific threat. Further details can be found in Section 564 of PAHPRA.

IV. OK, I get why and how the EUA was created. But how is an EUA issued?

There are five key components for an EUA issuance: determination of an emergency, declaration of an emergency, review of the EUA request by the FDA, approval and/or denial of the request and termination of the EUA.

Only the President can declare a federal emergency. Once a federal emergency is determined and declared by the President, one of three federal departments, the Department of Defense (DoD), the Department of Homeland Security (DHS), and the Department of Health and Human Services (HHS), can ask for an EUA. The DHS Secretary, him- and/or herself, can also ask for an EUA. The EUA is then reviewed by the HHS Secretary, who can justify that circumstances exist for the issuance of an EUA. Once this declaration is made, the FDA Commissioner then consults the CDC, the NIH and the HHS Assistant Secretary for Preparedness and Response (ASPR); based on these consults and the currently available evidence, the FDA can then issue the EUA. Of note, one of the key issues the FDA focuses on when deciding if a medical product should be issued an EUA is if the known and potential benefits outweigh the known and potential risks and/or side effects.

Because of this last point, most medical products, especially drugs and vaccines, must have gone through Phase 1 and 2 clinical trials, which prove safety and effectiveness, respectively. With vaccines, such as that of the Moderna, Pfizer and Johnson & Johnson, the FDA also expects data from an interim or finalized Phase III clinical trial with at least 3,000 participants; half of these participants must also have a median follow-up of at least 2 months. This follow-up period must include an analysis/description of serious adverse events and adverse events of interest of one-month duration.

Final Thoughts

This chapter was meant to provide further insight into what an EUA is, why it was created, its history, and the process of an EUA issuance. Some examples from the COVID-19 pandemic were used to provide further insight into the questions and confusion concerning the EUA. Hopefully, the reader will have a clearer understanding of what it means when a medical product has been given an EUA issuance.

¹ Commissioner, O. (2021, March 18). Emergency use authorization. Retrieved March 5, 2021, from <https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization>.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

References

1. Center for Biologics Evaluation and Research. (n.d.). Emergency use authorization for vaccines explained. Retrieved March 6, 2021, from <https://www.fda.gov/vaccines-blood-biologics/vaccines/emergency-use-authorization-vaccines-explained>.
2. Commissioner, O. (2021, March 18). Emergency use authorization. Retrieved March 5, 2021, from <https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization>.
3. Commissioner, O. (n.d.). EUA flow chart. Retrieved March 2, 2021, from <https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/summary-process-eua-issuance>.
4. Institute of Medicine (US) Forum on Medical and Public Health Preparedness for Catastrophic Events. Medical Countermeasures Dispensing: Emergency Use Authorization and the Postal Model, Workshop Summary. Washington (DC): National Academies Press (US); 2010. Emergency Use Authorization. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK53122/>.
5. Rizk, J. G., Forthal, D. N., Kalantar-Zadeh, K., Mehra, M. R., Lavie, C. J., Rizk, Y., Pfeiffer, J. P., & Lewin, J. C. (2021). Expanded Access Programs, compassionate drug use, and Emergency Use Authorizations during the COVID-19 pandemic. *Drug discovery today*, 26(2), 593–603. <https://doi.org/10.1016/j.drudis.2020.11.025>.

Submitted 4/2021

V.26 Integrating Evidence-Based Medicine into Journal Club (Simrun Bal)

“Medicine is a science of uncertainty and an art of probability.” – Sir William Osler (1)

As demonstrated in the familiar quote above by Sir William Osler, clinical medicine is complex, encompassing diverse scientific, mathematic, and humanistic skills and inherently involving varying degrees of uncertainty. Honing one’s craft in medicine involves learning to face uncertainty and grappling with probability, as Osler described, and the practice of evidence-based medicine (EBM) offers a complex set of skills to learn to do so in a thoughtful, careful and accountable manner. Journal clubs are one particular aspect of medical training programs that allow residents to learn and practice EBM principles pertaining to the clinical care of patients.

Journal clubs are ubiquitous in medical training programs. Specifically, they provide opportunities to help residents and faculty build skills in the critical assessment of medical literature, epidemiology, statistics, and research design (2).

This chapter provides a guide to how contemporary journal clubs in internal medicine training programs can benefit from a structured approach to journal club rooted in the principles of evidence-based medicine. The goal for residents and student learners is to optimally learn from devoted “journal club” time so that they can learn ways to examine and discuss evidence with the goal of making thoughtful decisions about clinical care and providing skilled communication and counseling to patients based on appropriate evidence.

How did journal clubs arise? As a brief historical review, the first mention of a “journal club” was from the memoirs of Sir James Paget, an English surgeon who described what became known as Paget’s disease of the breast and of the bones. He described that “some of the self-elect of the pupils, making themselves into a kind of club, had a small room [outside St. Bartholomew’s Hospital in London]...where we could sit and read the journals” (3). Later, Sir William Osler created the first formal journal club while at McGill University in 1875, with the aim of collectively reading subscription journals to help with learning while also reducing the high cost of print periodicals (4). By the early 20th century, most medical specialties at Johns Hopkins Hospital were hosting specialty-specific journal clubs, often in the homes of participating physicians (4). In contemporary times, journal clubs meet often in-person or virtually (sometimes nowadays even via social media platforms) with a resident or fellow presenting an article and with a chief resident or faculty member guiding discussion and facilitating participation. Faculty members offer expert opinions and stimulate debate about certain key points.

Given the strong history of journal clubs throughout history and the contemporary role these groups play in the critical appraisal of medical literature and the dissemination of new evidence, it is helpful to consider a structured approach for learners to integrate principles of evidence-based medicine into journal clubs.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

The following approach is suggested based on review of the Evidence Based Medicine Elective (taught by Dr. Ross), the textbook *Evidence-Based Medicine* (5), curricular review of the FIRE rotation (Formal Instruction in Resident Education), and a suggested guide by Dr. Thomas Newman, MD, MPH, a professor of clinical epidemiology at the University of California at San Francisco (6).

Structured Approach to Integrating Evidence-Based Medicine into Journal Club

Select an appropriate article

Selecting an appropriate article begins with first asking a good question, which is at the heart of evidence-based medicine and the clinical care of patients as well. Think back on your experiences in the wards and asking questions of specialty colleagues. When a clinical question is clearly articulated, it is more likely to lead to receiving clear and timely answers (5).

For journal clubs involving interns/residents or fellows (rather than individuals earlier in training), it is helpful to build a question that would directly inform a “foreground” (rather than “background”) clinical decision that would be made with a patient. Foreground questions are focused on specific knowledge that would inform a clinical decision. They have four main components: (taken from p. 21, Reference 5).

1. The patient situation, population, or problem of interest
2. The main intervention (such an exposure, diagnostic test, prognostic factor, treatment, patient perception, etc.)
3. A comparison intervention or exposure, if relevant
4. The clinical outcome of interest, including a time horizon

This framework is the basis for the PICO structure (**p**opulation/**p**roblem, **i**ntervention, **c**omparison/**c**ontrol, and **o**utcome), which is described later in this chapter and will inform part of your presentation.

Once you have formed a question, the next step is finding an article. The article should ideally report original research, rather than being a review article, as it is crucial to analyze the methods section of the paper.

Prepare the participants and outline expectations

It is helpful to clarify the goals of the journal club, which vary depending on the stage of training. At Dartmouth-Hitchcock, the Internal Medicine Residency Program has a journal club every other week hosted by the resident who is participating in the “FIRE” (Formal Instruction in Resident Education) Rotation. This resident is responsible for asking the clinical question based on a clinical encounter, finding an article to address the question, and performing a critical appraisal of the article through the Journal Club.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

As you prepare, it is important to clarify expectations: do you want participants to have to read the article? If so, aim to distribute the article via email about 7 days in advance. Communicate that you are looking forward to resident participation and plan to utilize verbal and nonverbal techniques to encourage all to participate. Bring extra copies of the article to the in-person session for those who may not have read the article.

Review the article yourself

Take your time reading the study in a careful and critical manner, utilizing the format below. For those who have not taken the Evidence Based Medicine Elective, it would be helpful to review some of the basics of EBM (see <http://ebm.harley.ninja>). The EBM Guide on this site also offers a variety of chapters describing study design, research methods and statistics, and the critical appraisal of evidence, which can be helpful as you prepare your presentation.

As you critically read the study, try to determine a few main points or concepts that you find important in reading. For example: hazard ratio, effect size, confidence intervals, historical aspects of race in research studies, and more. Refer to the EBM Guide to further your understanding of these concepts and pick out 1-2 concepts that you would like to discuss in your presentation (this will be discussed later in the framework).

Prepare to lead the discussion

Before your presentation (see below structure), remind yourself of a few basic principles that make the experience of Journal Club more enjoyable and interactive for your colleagues.

1. Timing: Remember to time yourself in advance (typically journal clubs will take about one hour). Starting and ending on time is crucial. A suggested time-based approach is to take a few minutes with describing why you chose the article, a clinical vignette, and background information, then take about 20 minutes describing the “objective” aspects of the study (described below), and an additional 25-30 minutes focused on an interactive discussion highlighting more “subjective” aspects of the design, analysis, and the implications of the study on clinical practice.
2. Interactivity: One of the gifts of Journal Club is the diversity of background of the participants. To make the presentation interactive, try to avoid answering the questions that you pose, and ask residents and faculty members to offer their interpretations. Learn from the perspectives of different faculty members who may have much more experience in research.

Present key points from the article in the following format

1. Title/theme of today’s journal club
2. Learning objectives
 - a. Briefly outline the main objectives that you hope to accomplish in your presentation
3. Case vignette / Clinical problem

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

patients in the study, and any areas of disparity that you observe. At this stage, having this information will be helpful later when you determine internal validity and external validity.

- b.** Introduce the group to the intervention and the control (briefly), then discuss how patients were randomized, as well as the duration of the study and the follow-up period.

9. PICO Framework: Describe the intervention

- a.** Describe what happened to the intervention group. If the trial was a randomized control trial, discuss if the treatments were blinded for participants, providers, and/or researchers.

10. PICO Framework: Describe the control

- a.** Describe what happened (instead of the intervention) to the control group

11. PICO Framework: Describe the outcomes

- a.** Describe the primary outcome and any secondary outcomes that the trial defined and consider how these outcomes were measured.
- b.** Using statistics, describe the major outcomes from the study and how the data was analyzed. Consider not only the statistical significance of the outcomes, but consider the effect size as well, in terms of understanding the magnitude of the difference between groups. Try to report statistics in terms of EER (experimental event rate) and CER (control event rate), and from there, calculate the relative risk reduction (RRR) and the number needed to treat (NNT) or number needed to harm (NNH). These principles are explained in the EBM Guide. Utilize the EBM calculator available at <https://ebm-tools.knowledgetranslation.net/calculator> to convey the effect sizes.
- c.** For a randomized controlled trial, it is helpful to consider if the analysis was performed by an “intention to treat” analysis or “as treated” (analysis by treatment received). Consider also the drop-out rate as well as the completeness of follow-up, as well as the follow up time period.
- d.** Often, major results are summarized in tables or figures. It can be helpful to review the most important tables or figures in your presentation so that everyone understands the results. Magnify all tables or figures so that one slide has approximately one table/figure at maximum size.
- e.** In this section, you may discuss statistical concepts that serve as learning opportunities for both the presenter of journal club as well as the audience. Take your time to discuss the 1-2 key statistical concepts that you made note of when

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

reading the paper earlier, making use of the rest of the EBM Guide to understand and teach about these concepts.

- f. Discuss the conclusions that the authors draw from the results (don't discuss yet if you agree with the conclusions).
- g. In this section, it is helpful to also discuss adverse events. Presenting both the NNH (number needed to harm) and the NNT (number needed to treat) allows you to describe the balance between a potentially harmful outcome vs a potentially beneficial outcome.

12. Discuss the validity of the study

- a. At this point, you as the facilitator will transition from discussing the "objective" aspects of the paper to providing an interpretation of the paper.
- b. Start by identifying possible biases or flaws in the study, considering the appropriateness of the study design and the validity of measurements. One can consider the analyses of possible biases as similar to developing a broad differential diagnosis. Consider the effect of bias on the overall validity of the results.
- c. Comment specifically on the high-quality aspects of the study (study strengths) as well as weaknesses (areas where you feel the study could improve). In considering strengths, pay particular attention to the study design, population, and methodology, and reflect back on your clinical question to determine if the study was appropriate in answering the initial clinical question. In considering weaknesses, it is important to reflect on the potential sources of bias that were identified when reviewing study design and outcomes.

13. Formulating a conclusion and fostering reflection

- a. Formulating a conclusion is a crucial part of the presentation, as it is where you as the facilitator, paired with the group, not only work together to cultivate a conclusion together utilizing both objective and subjective analysis, but also where the group begins to question how they may integrate the findings from the paper into their clinical practice. This step involves you initially describing your own conclusions regarding the findings, integrating thoughts on validity and quality, to create an understanding of whether the results were clinically meaningful and if the study was scientifically sound. Importantly, your conclusion may differ from that of the authors, and it is helpful to discuss the difference in conclusions (and the clinical and statistical reasoning behind the difference).
- b. In this section, it may be helpful to return to the original clinical vignette to describe if one's clinical practice would be changed due to the results of the study and to contemplate if the study would affect one's care of the patient(s) who originally sparked the clinical research question. In considering these questions, also

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

contemplate if there are remaining gaps that linger in regard to our understanding of the question and problem. Lastly, to spark discussion and reflection among the group, ask group members how they might design a “next study” to address the gaps that are identified. You could also consider posing a clinical scenario to the group and asking about potential management strategies based on the findings of the paper.

Personal contemplation and growth

After your journal club is finished, take a few moments to reflect upon how the session went. Think about areas where you noticed yourself struggling with the articulation of different statistical concepts and make note of these. It will be helpful to review these concepts again in the EBM Guide as well as in the textbook *Evidence-Based Medicine* (Straus, et al; Reference 5). Contemplate times when you felt that the cohort was participating actively and in response to which questions that were posed or teaching methods that you utilized. Remember that the process of both learning and teaching evidence-based medicine is one built on deliberate practice and reflection. It may be helpful to contact a chief medical resident or a faculty member within the Department of Medicine to provide feedback on your teaching skills.

References:

1. The Lancet. Editorial: Uncertainty in medicine. *The Lancet* 2010; 375 (9727): 1666. [https://doi.org/10.1016/S0140-6736\(10\)60719-2](https://doi.org/10.1016/S0140-6736(10)60719-2).
2. Alguire PC. A review of journal clubs in postgraduate medical education. *J Gen Intern Med*. 1998 May;13(5):347-53. doi: 10.1046/j.1525-1497.1998.00102.x. PMID: 9613892; PMCID: PMC1496950.
3. Linzer M. The journal club and medical education: over one hundred years of unrecorded history. *Postgrad Med J* 1987; 63(740): 475-478.
4. Topf JM, Sparks MA, Phelan PJ, et al. The evolution of journal club: from Osler to Twitter. *AJKD* 2017; 69(6): 827-836.
5. Straus SE, Glasziou P, Richardson WS, and Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM*. 5th edition. Elsevier; 2019.
6. Newman, T. Suggestions for leading a journal club. June 11, 2007. Accessed December 14, 2021. https://libguides.library.arizona.edu/ld.php?content_id=58432543

Submitted 12/2021

Section VI. Integrating Diverse Sources of Information

VI.1 A Dive into Diabetes Management (Patrick Puliti)

The Evidence Based Medicine Elective enhances one's skills in accessing, interpreting, evaluating (quality and validity) and summarizing high quality research evidence that has the potential to change practice or support existing practice. It also enhances communication of evidence with both peers and patients when one explores the methods and outcomes of such studies. In reviewing a great deal of information regarding diabetes management, the possibility of integrating multiple sources of information to address a common clinical challenge led me to summarize such information this review.

It has long been the standard of care that metformin is the first choice for the majority of patients diagnosed with Type 2 Diabetes. [1]. With increasing development of newer anti-diabetic

medications there has been a shift away from early utilization of insulin and sulfonylureas to the newer second-line medications following failure of glycemic control with metformin alone. In the *Standards of Medical Care in Diabetes – 2021*, new recommendations that DPP-4 inhibitors/GLP-1 agonists, SGLT-2 inhibitors or thiazolidinediones should be initiated following metformin inadequacy for achieving glycemic control. (Figure 1) It should be noted that these standards are an ideal and, with many newer medications, are dependent on adequate coverage by insurance providers. What is the evidence behind their usage that makes them more beneficial than insulin? We will cover these four medication classes and what evidence supports their use below, as well as reference a few head-to-head trials that have been performed. Note that this is far from an exhaustive list of trials, but instead, a sampling of the more notable published trials that are of high quality.

GLP-1 Agonists

GLP-1, standing for Glucagon-like Peptide 1, is an incretin that acts on the GLP-1 receptor. Activation of this receptor results in stimulation of insulin synthesis and secretion, as well as slowing of gastric emptying and inhibiting post-meal glucagon release. Of note, native GLP-1 is typically only stimulated by oral glucose intake. GLP-1 is a short acting molecule that undergoes degradation by DPP-4, which is another therapeutic target that is described in the next section. GLP-1 agonists are a class of medications that seek to stimulate the GLP-1 receptor longer than native GLP-1 and are resistant to degradation by DPP-4. [2] Because of their activity in slowing gastric emptying, they have also been observed to result in decreased food intake and subsequent weight loss, with some medications in this class being FDA approved for weight loss in patients who do not have diabetes.

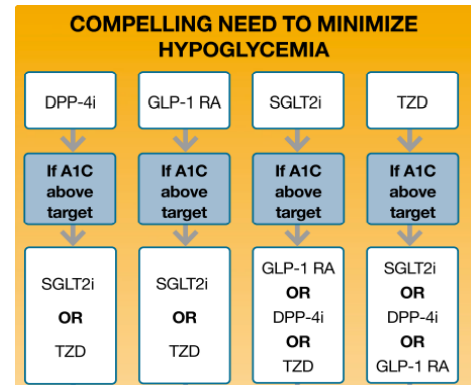


Figure VI-1: *Standards of Medical Care in Diabetes*, medications to start after Metformin [1]

What are the benefits of GLP-1 agonists? For starters, let's take a look at the evidence that supports its use in diabetes as an agent that can improve glycemic control.

The PIONEER 1 trial looked at the use of oral semaglutide in 3mg, 7mg and 14mg dosages, compared to placebo, in patients with Type 2 Diabetes over the course of 26 weeks. The average initial Hemoglobin A1c was 8.0% in this population. A significant improvement in Hemoglobin A1c was seen in all three treatment groups with increasing degree of improvement with increased dose (Figure 2). The primary endpoint of this study was the percent of patients that achieved a HgbA1c of less than 7%. Each dose was significant in its improvement over placebo, with 55.1% in the 3mg daily group, 68.8% in the 7mg daily group, and 76.9% in the 14mg daily group, with 31% of patients receiving placebo achieving a HgbA1c less than 7. Although not a primary endpoint, weight loss has been previously reported with GLP-1 agonists and this was measured as well. However, there was not a significant difference between Placebo and either the 3mg or 7mg daily groups, but there was significant change seen when placebo was compared to 14mg daily, with the average weight loss of 3.7kg in the 14mg daily group compared to 1.4kg in the placebo group [Figure 3]. [3]

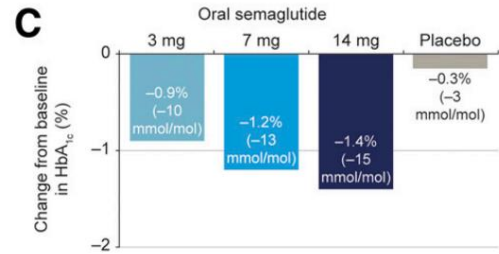


Figure 2: Change in A1c after 26 weeks of

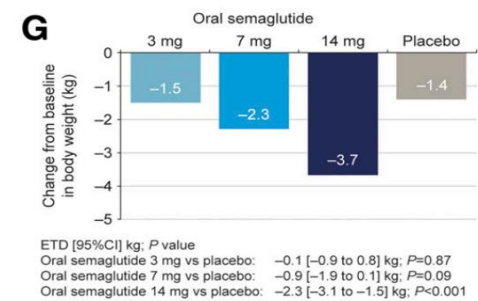


Figure 3: Change from baseline body weight (kg) after 26 weeks of semaglutide treatment

What about GLP-1 agonists and Cardiovascular and Renal Outcomes?

Multiple randomized control trials have sought to establish the benefit of GLP-1 agonists in regards to cardiovascular outcomes. The LEADER trial recruited patients with diabetes, age greater than 50, who had either coronary artery disease, peripheral vascular disease, CKD stage 3, or NYHA Class II/III heart failure. Subjects were initiated on either subcutaneous Liraglutide 1.8mg daily or Placebo. The primary endpoint after a mean of 3.8 years of follow-up was a commonly chosen composite outcome of “Death from Cardiovascular Causes, non-fatal Myocardial Infarction, or non-fatal Stroke”. Compared to placebo, there was a 12% relative risk reduction with a number needed to treat of 54. A follow-up to the LEADER trial looked at the data to assess for progression of renal disease with a composite outcome of macroalbuminuria, doubling of Creatinine, progression to ESRD, or death from renal causes. Amongst the treatment group with liraglutide compared to placebo, there was a relative risk reduction of 20% for this composite, and a NNT of 68.

The REWIND trial looked at subcutaneous Dulaglutide 1.5mg weekly in patients with Type 2 Diabetes. Subjects were 50 years or older with HgbA1c less than 9.5%. The trial included those older than 50 with known vascular disease, patients older than 55 (with history of myocardial infarction, coronary/carotid/lower extremity stenosis greater than 50%, Left Ventricular Hypertrophy, or eGFR less than 60) or men older than 60 years with two of the following: Tobacco use, dyslipidemia, hypertension, or abdominal obesity. A similar composite primary outcome of non-fatal MI, non-fatal stroke, or death from cardiovascular causes was measured over a median follow-up of 5.4 years. For the treatment group on Dulaglutide compared to placebo, there was a 10% relative risk reduction with a NNT of 71, a remarkably similar efficacy outcome. [5]

Lastly, the SUSTAIN-6 studied patients with diabetes age 50 or older (with at least one of: heart failure, CKD stage 3 or above, or established CV disease) or patients with diabetes age 60 or older with at least one of the following: persistent microalbuminuria or proteinuria, LVH on EKG or echo, LV dysfunction on imaging, ABI < 0.9. Similar to the previous trials, the composite outcome was death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke. Compared to placebo, there was a 26% relative risk reduction with a number needed to treat of 43. [6] Thus, GLP-1 agonists of different agents have a similar impact on serious. Outcomes, with NNT range of 43-71 to avoid serious outcomes.

DPP-4 Inhibitors

Functioning similarly to GLP-1 agonists, DPP-4 inhibitors seek to block the DPP-4 enzyme that is responsible for the degradation of GLP-1, resulting in a similar end-result of allowing GLP-1 to activate GLP-1 receptors (stimulating insulin synthesis/secretions, slowing gastric emptying, and inhibiting post-meal glucagon release [7]). Unlike GLP-1 agonists, which until the recent release of oral semaglutide were primarily subcutaneous medications, DPP-4 inhibitors are primarily oral medications. Although they function well in glycemic control, their cardiovascular outcomes are not as strongly supported by randomized control trials.

The VERIFY trial looked at patients with a HgbA1c between 6.5-7.0% and treated them with either Vidagliptin with Metformin or Placebo with Metformin. The trial sought to evaluate the risk of treatment failure, defined as a hemoglobin A1c greater than 7.0% for two consecutive visits, with visits scheduled at 13 week intervals, over the course of 5 years. Amongst the group treated with Vidagliptin/Metformin compared to Placebo/Metformin, there was a 30% reduction in treatment failure and a number needed to treat of 5. [8] The use of a surrogate endpoint, A1c, rather than a clinical outcome, makes comparisons with the previous class of medications difficult.

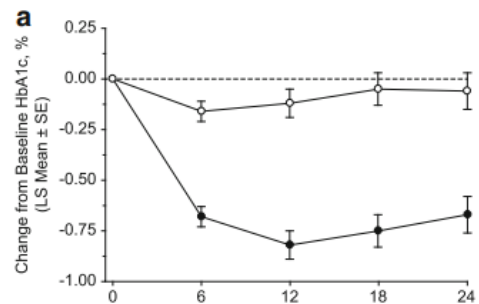


Figure VI-2: Change in Hemoglobin A1c after 24 weeks of treatment of Omiglipitin compared to placebo [9]

A separate trial looked at Omarigliptin 25mg weekly in patients with an A1c 7.5-10.5% on Metformin and glimeperide. Although this was standard of care at that time and not a fault of the study, the generalizability of this study is reduced with the newer recommendations of DPP-4 inhibitor initiation prior to sulfonylurea usage, as per the aforementioned *Standards of Medical Care in Diabetes*. Nonetheless, there was a significant improvement in Hemoglobin A1c, with an average reduction of 0.61% after 24 weeks for the Omarigliptin group compared to placebo [Figure 4] [9]. Again, surrogate endpoints rather than clinical outcomes are problematic in assessing clinical efficacy.

Was there an improvement in Non-Alcoholic Steatohepatitis?

There are only a few trials that are available concerning patients with diabetes and NASH treated with DPP-4 inhibitors. One trial sought to look for histological improvement of NASH via liver biopsy, the gold-standard. The trial included 12 patients with biopsy proven NASH, and placed 6 of them of Sitagliptin and 6 on placebo for a total of 24 weeks; investigators measured liver steatosis on repeat biopsy and hepatic fat fraction on MRI. Unfortunately, there was no significant difference in Sitagliptin compared to Placebo in either of these outcomes. Interestingly, the study did not detect a significant change in Hemoglobin A1c amongst the two groups, possibly an indicator of the small numbers of participants and low power of the study. [10]

Do DPP-4 inhibitors have similar evidence as GLP-1 Agonists in cardiovascular or renal outcomes?

Although there have been a few trials that have sought to establish DPP-4 inhibitors as having similar benefits as GLP-1 agonists, these trials have not been able to establish improvement in cardiovascular outcomes. In the CARMELINA trial, patients with diabetes (A1c 6.5-10.5%) and an elevated cardiovascular risk (known coronary artery disease, stroke, or peripheral vascular disease) or elevated risk of renal disease (eGFR 45-75 with an elevated Urine Albumin: Creatinine ratio or an eGFR 15-45) were treated with either Linagliptin 5mg daily or Placebo. The primary outcome was a composite of time to first occurrence of cardiovascular death, non-fatal myocardial infarction, or non-fatal stroke, with a secondary outcome as time to end-stage renal disease diagnosis, reduction in eGFR of at least 40%, or death due to renal failure [Figure 5]. Unfortunately, there was no significant difference in either of

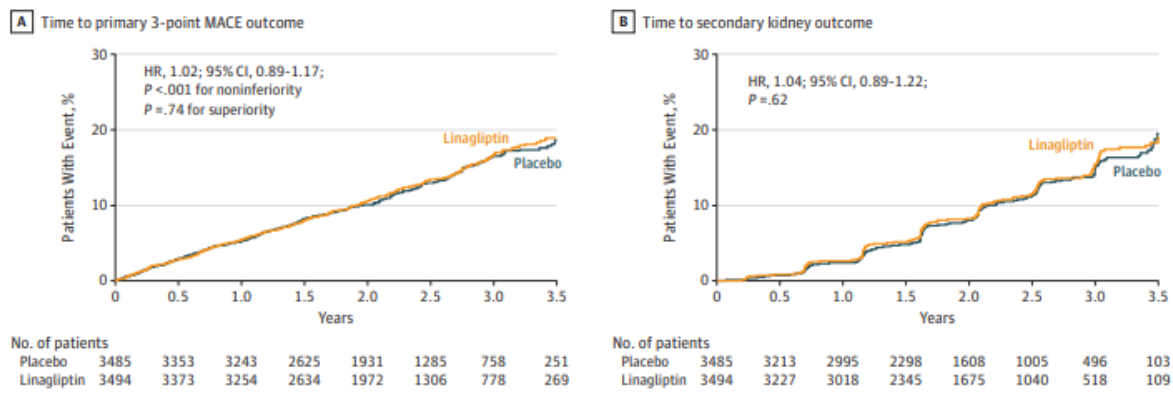


Figure VI-3: Time to 3-point MACE outcome and time to renal outcomes in Linagliptin compared to placebo [11]

these outcomes for linagliptin compared to placebo. [11]

The SAVOR-TIMI 53 trial looked at patients with Type 2 Diabetes with A1c 6.5-12% and 40 years of age who had previously had a clinical event of coronary, cerebrovascular, or peripheral vascular disease, or age at least 55 (men) or 60 (women) with dyslipidemia, HTN, or active smoking and treated them with Saxagliptin 5mg daily or placebo. Primary composite outcome was cardiovascular death, non-fatal myocardial infarction, or non-fatal ischemic stroke, with a secondary composite outcome of heart failure, unstable angina, or coronary revascularization. In both of these outcomes, there was no significant difference between placebo and Saxagliptin [Figure 6]. Of note, when hospitalization for heart failure was isolated, there was a relative risk increase in the Saxagliptin treatment group of 11% with a number-needed-to-harm of 18, though this has not been observed in other studies. [12]

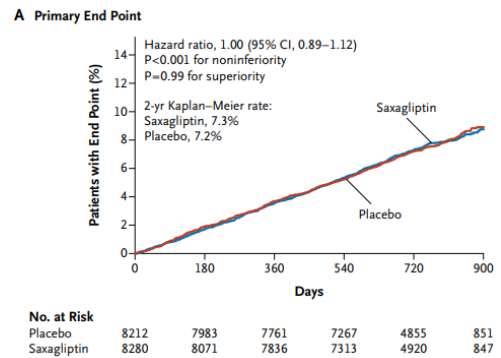


Figure VI-4: Composite of Cardiovascular death, nonfatal MI, or nonfatal ischemic stroke in Saxagliptin compared to Placebo [12]

Thus, one is led to believe that there is an insufficient evidence base on which to recommend use of DPP-4 inhibitors at this point.

SGLT-2 Inhibitors

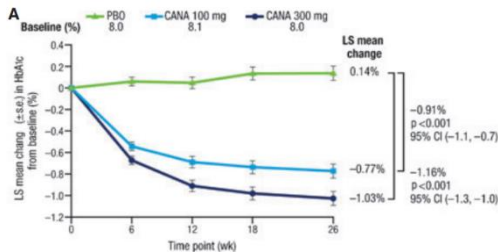


Figure 7: Change in Hemoglobin A1c amongst patient's treated with Canagliflozin 100mg daily, 300mg daily, or placebo [14]

SGLT-2 inhibitors inhibit the Sodium-Glucose Transporter-2, which is located in the proximal convoluted tubule and is responsible for 90% of the renal resorption of filtered glucose. Over the course of 24 hours of inhibition of the SGLT-2 transporter, 60-80g of glucose are excreted and not resorbed, which calculates to about 240-320 excess calories excreted in the urine. [13]

In regard to its efficacy in glycemic control, patients with diabetes (A1c 7-10%) were treated with either Canagliflozin 100mg daily, 300mg daily, or placebo over 26 weeks. The primary outcome of the study was the percent of patients that were able to achieve a hemoglobin A1c less than 7.0%. Amongst those in the treatment group, there was a significant improvement in the primary outcome with 62.4% of patients on 300mg daily, 44.5% of patients on 100mg daily, and 20.6% of patients on placebo reaching the primary endpoint. Additionally, a secondary outcome was the percent reduction in A1c after 26 weeks, which was also significantly different, with a 1.03% reduction in A1c with canagliflozin 300mg daily compared to 0.14% increase in A1c with placebo [Figure 7] [14]. Again, the investigators chose a surrogate and not a clinical outcome, which undoubtedly is easier to show over the course of a short study.

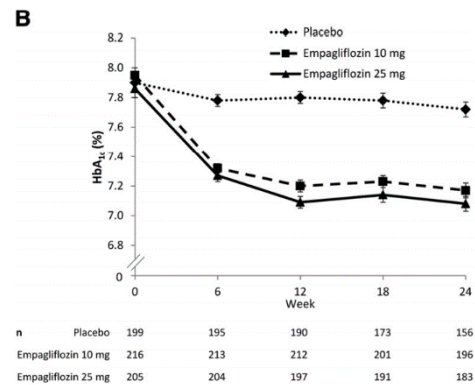


Figure 8: Change in HgbA1c after 24 weeks of treatment with Empagliflozin compared to placebo [15]

Another study looked at Empagliflozin treatment in 638 patients with diabetes with Hemoglobin A1c between 7.0-10.0% despite a diet/exercise program and treatment with metformin greater than 1500mg per day. These patients were treated with either Empagliflozin 10mg daily, 25mg daily, or placebo. The endpoint for the study was looking at change in baseline Hemoglobin A1c at the 24 week mark, with secondary endpoints in change in body weight. Both treatment doses had significant decreases in Hemoglobin A1c compared to placebo, with a -0.57% absolute decrease in the 10mg daily group, and a -0.64% absolute decrease in the 25mg group [Figure 8]. For the secondary outcome of weight loss, the placebo group had a 0.45kg reduction in body weight in both treatment groups with a 2.08kg reduction in the 10mg daily group and 2.46kg in the 25mg daily group, a result that was significant. [15]

In patients with heart failure, were SGLT-2 inhibitors beneficial?

In the SOLOIST-WHF trial, patients with diabetes and a diagnosis of heart failure (median ejection fraction of 35%) that were recently admitted for decompensated heart failure were started on Sotagliflozin 200mg (up-titrated to 400mg daily) with follow-up over 9 months. The primary endpoint was a composite of total events including death from cardiovascular causes, or hospitalizations/urgent visits for heart failure. There was a significant reduction in the treatment group with an HR of 0.67 [Figure 9]. Additionally, there was a 37% relative risk reduction with a NNT of 4 in the treatment group for hospitalizations or urgent visits for heart failure. There was a significant increase in two adverse events, genital mycotic infections occurred in 2.4% of the treatment group compared to 0.9% in placebo, and diarrhea was increased 8.5% in the treatment group compared to 6.0% in the placebo group [16]

In the CANVAS trial, patients with Type 2 diabetes (A1c between 7-10.5%), age 30 years or older with either a history of atherosclerotic cardiovascular disease or age at least 50 years with two or more: 10 or more years of diabetes, Systolic blood pressure higher than 140mm Hg on one or more anti-hypertensives, smoking, HDL <38, or macro/microalbuminuria, were treated Canagliflozin 300mg, 100mg, or placebo. In the Canagliflozin 300mg group, there was 37% relative risk reduction in heart failure admissions for an NNT of 114, and a 16% relative risk reduction in deaths due to cardiovascular causes, non-fatal myocardial infarction, and non-fatal stroke for an NNT of 68 [Figure 10]. Recall that this effect size is similar to that demonstrated for GLP-1 agonists. During this study however, there was a significant increase in genitourinary tract infections for those receiving the treatment, with 68.8 compared to 17.5 events per 1000 patient years for women and 34.9 compared to 10.8 events per 1000 patient years for men. There was also a significant increase in amputations in the treatment group with 6.3 compared to 3.4 events per 1000 patient years. [17]

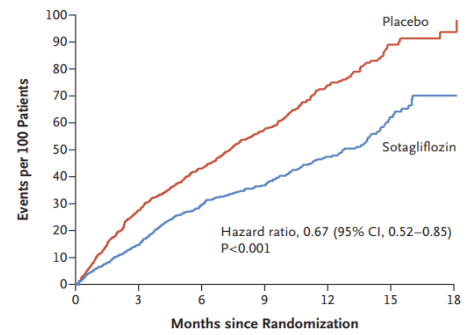


Figure 9: Total events (deaths from CV causes, or Hospitalizations/Urgent visits for heart failure) in Sotagliflozin compared to Placebo [16]

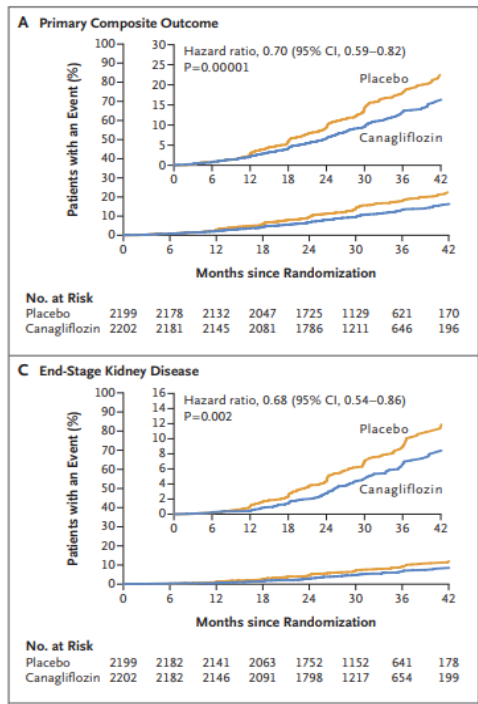


Figure 10: Treatment with canagliflozin versus placebo in patients with Type 2 Diabetes and CKD [17]

In patients with diabetes and CKD were SGLT-2 Inhibitors beneficial in preventing progression?

The CREDENCE trial looked at patients with A1c 6.5-12% and a diagnosis of CKD with an eGFR 30-90 with albuminuria and treatment with ACE-I/ARB that were randomized to Canagliflozin 100mg daily or Placebo over 2.5 years. A composite outcome was progression to end-stage renal disease, doubling of creatinine, or renal/cardiovascular death. For this composite outcome, there was a 28% relative risk reduction with a NNT of 23. When isolated for progression to End Stage Renal Disease, there was a 30% relative risk reduction with a NNT of 45. Although previous trials have noted the cardiovascular benefits of SGLT-2 inhibitors, there was no difference in cardiovascular deaths between placebo and Canagliflozin treatment in this study. [18]

Thiazolidinediones (glitazones)

Thiazolidinediones, often abbreviated TZDs, are selective peroxisome proliferator-activated receptor-gamma (PPAR-gamma) agonists. They function by increasing the insulin sensitivity of muscle and adipose tissue [19]. In a randomized control trial looking at their efficacy in glycemic control, 408 patients with A1c greater than 7% (average 10.4%) without a diagnosis of neuropathy, impaired liver function, or a history of MI/CABG/TIA/CVA were randomized. Patients were treated with pioglitazone versus placebo over the course of six months. Those treated with Pioglitazone initially 15mg daily and were then uptitrated to 45mg daily. After six months, there was a statistically significant decrease in A1c in the treatment group, with a mean reduction in A1c of 1.6%. [20]

What are the cardiovascular benefits of TZDs?

The ProACTIVE trial sought to evaluate the cardiovascular benefits of Pioglitazone in the treatment of diabetes. Patients with diabetes between 35-75 years old not on insulin with evidence of macrovascular disease, defined as myocardial infarction or PCI/CABG at least 6 months before, or diagnosis of ACS at least 3 months prior, were treated with Pioglitazone or placebo. The composite outcome was time to first occurrence of: all-cause mortality; nonfatal MI; acute coronary syndrome; cardiac intervention, including CABG or PCI; stroke; major leg amputation or revascularization in the leg. There was no significant difference in the composite outcome between the placebo and the treatment group.

Unfortunately, there were notable multiple adverse events that were significant in the Pioglitazone group. There was a significant increase in the reports of overall heart failure events with a relative risk increase of 43.4% and a NNH of 31. Additionally, although they were not statistically significant there was a concern for an increased risk of bladder cancer in the treatment group with 14 cases of bladder cancer in the treatment group compared to 6 in the placebo. An unaffiliated panel of experts evaluated the cases and noted that 11 of the total of 20 cases could not be attributed to the treatment. Of the six remaining cases in the treatment group, 4 had a history of tobacco use, and it was felt by the panel that the overall increased cases could not be attributed to Pioglitazone. [21]

The risk of heart failure exacerbation became an increasing concern with the use of TZDs given the overlapping populations of patients with both diabetes and heart failure. Lincoff et al. performed a meta-analysis of randomized control trials with Pioglitazone treatment (n = 8554) versus Control (n = 7836). Although the outcome of “Serious Heart Failure” was not explicit, there was a significantly increased risk of heart failure exacerbation in the Pioglitazone population, with a calculated hazard ratio of 1.41. [22]

Is there improvement in NASH?

A study focusing on the histological improvement of NASH in patients with diabetes or pre-diabetes treated with Pioglitazone was performed. Patients were treated with Pioglitazone 30mg (with an increase to 45mg after two weeks) or placebo. Prior to randomization, patients NASH was diagnosed histologically by biopsy. After 18 months of treatment, biopsy was performed again. The primary outcome was the Non-Alcoholic Fatty Liver Disease Score (NAS), which is a histologic scoring system: out of a total score of 8 with a 5-8 considered a diagnosis of NASH. Investigators considered the primary outcome to be an at least 2 point reduction in the score, with a secondary outcome being resolution of NASH, which is a score less than 5. There was a 241% relative improvement in the treatment group compared to placebo in those who had an at least 2 point reduction for a NNT of 2. Additionally, there was a 168% relative improvement in the treatment group for resolution of NASH compared to placebo, for a NNT of 3. [23]

Are there Head to Head Trials of these four medication classes?

There are an extensive literature of head to head trials for these medications, particularly amongst the newer SGLT-2 inhibitors, DPP-4 inhibitors, and GLP-1 agonists. Outside of the similar mechanisms for GLP-1 and DPP-4 making them mutually exclusive, the remaining medications are intended to be additive, and these trials allow practitioners to prioritize one class over another, depending on the patient. Below are a few trials that have been conducted:

Semaglutide compared to Empagliflozin

The Pioneer 2 Trial compared Semaglutide (GLP-1 agonist) and Empagliflozin (SGLT-1 inhibitor). Compared to Empagliflozin, Semaglutide had a statistically significant improvement in its Hemoglobin A1c reduction, with an average reduction of 0.4% compared to 0.9%, for a greater reduction in Hemoglobin A1c of 0.4% [Figure 11-Top]. Additionally, as weight loss is an intended side-effect of both medications, the study looked at the change in body weight over 1 year during the trial. Interestingly, despite only Semaglutide having approval for use as a weight loss medication outside of the setting of Diabetes, there was no significant difference in body weight change. The group treated with Semaglutide had -3.8kg weight change at both 26 and 52 weeks, and the Empagliflozin group had a -3.7kg weight change at 26 weeks, and -3.6kg weight change at 52 weeks [Figure 11-Bottom]. Although this trial showed a slightly better glycemic outcome, it provided further evidence that Empagliflozin should also be considered when weight loss is a goal in patients. [24]

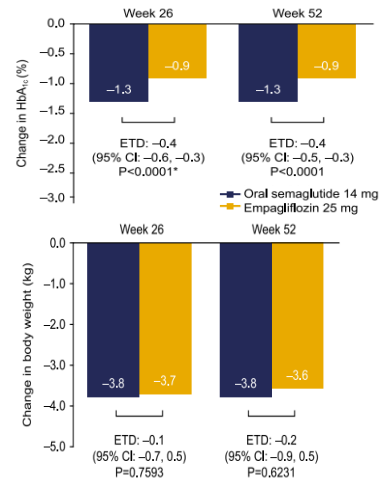


Figure 11: A1c improvement (top) and Body Weight Change (bottom) over 52 weeks in Semaglutide compared to Empagliflozin [24]

Oral semaglutide compared to subcutaneous liraglutide

Unlike the previous trial, the Pioneer 4 trial’s purpose was to compare one GLP-1 agonist compared to another. This trial was particularly interesting as GLP-1 agonists generally have been subcutaneous medications. Oral semaglutide has both a growing body of evidence to support its efficacy and it has the added benefit of not requiring subcutaneous injections, a likely barrier to both adherence and adoption of this class of medication. Although at 26 weeks the medications had no significant difference in Hemoglobin A1c reduction, after 52 weeks there was a significant difference, primarily due to subcutaneous Liraglutide’s A1c change going from -1.1% at 26 weeks to -0.9% at 52 weeks [Figure 12]. This study also looked at body weight change, and at both 26 and 52 weeks, Semaglutide (4.3kg reduction at 52 weeks) showed a significant improvement in body weight reduction compared to Liraglutide (3.0kg reduction at 52 weeks). [25]

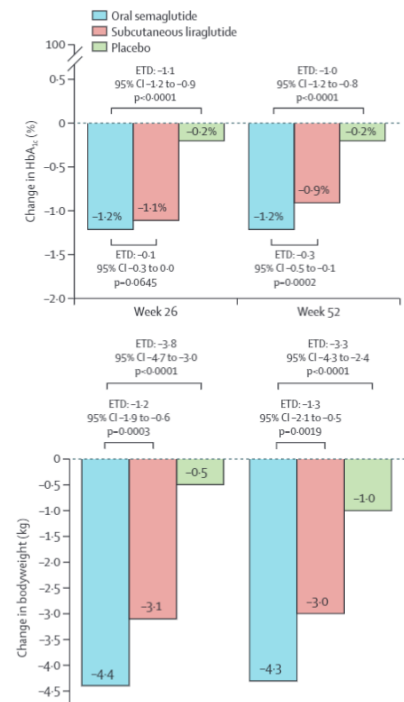


Figure 12: A1c improvement (top) and Body Weight Change (bottom) over 52 weeks in Semaglutide compared to Liraglutide [25]

Oral Sitagliptin compared to Subcutaneous Dulaglutide

DPP-4 inhibitors and GLP-1 agonists share a similar pathway that they are involved in, resulting in the medications being exclusive to one another. As discussed in the above sections, GLP-1 agonists appear superior to DPP-4 inhibitors in that they have significant improvement in cardiovascular, renal and weight loss outcomes that has not been similarly documented in DPP-4 inhibitors. With this in mind, the AWARD-5 trial compared Dulaglutide 1.5 and 0.75mg weekly to Sitagliptin over 2 years. Over the course of the study, there was significant difference in both dosages of Dulaglutide (0.99% A1c reduction for 1.5mg weekly and 0.71% A1c reduction for 0.75mg weekly) compared to Sitagliptin, which only had a 0.32% reduction in hemoglobin A1c [Figure 13]. [26]

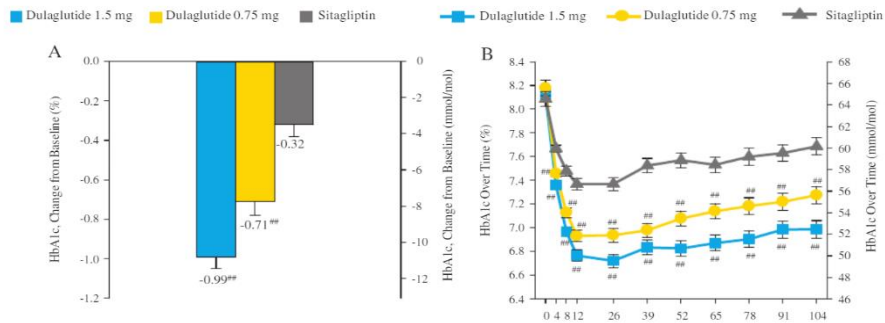


Figure 13: Change in A1c after 2 years (left) and over time (right) for Dulaglutide compared to Sitagliptin [26]

Conclusion

There is an immense amount of data surrounding diabetes management and this is just a brief look at 4 classes of medications that have become increasingly recommended as next steps in the management of diabetes after failure to achieve glycemic control with metformin. Each medication has improved glycemic control, with some of the trials showing greater reductions with some medications compared to others. Where these medications differ most, is their observed benefit in comorbid conditions that often accompany Type 2 Diabetes. Using cardiovascular death, non-fatal MI, and non-fatal stroke as a commonly utilized outcome measure, we can see that DPP-4 inhibitors and TZDs, although adequate for glycemic control, did not show a significant improvement in this outcome. Conversely, SGLT-1 inhibitors and GLP-1 agonists both had a NNT ranging from 43-71, suggesting that they are likely equivalent and can both be utilized for this benefit in patients with elevated risks. This is simply one example of where the evidence can be applied to guide the treatment options available, depending on the nuance of each patient and their comorbid conditions. Recommendations, such as the American Diabetes Association's *Standards of Medical Care in Diabetes*, provide a starting point for practitioners with evidence based next steps, however the nuance of when each medication has added benefit, comes through in the data. Hopefully, this chapter provides an introduction to some of this evidence and provides some resources in guiding the management of your next patient with a diagnosis of diabetes.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Beyond those conclusions, this exercise reinforces some principles that go well beyond diabetes and should guide treatment decisions in other disease states. For the inquiring clinician, and indeed the informed patient, some questions that inform one's decision-making might include the following:

- What is a significant NNT or NNH?
- In considering different medications, have investigators used similar patients, and similar outcomes?
- Are surrogate outcomes significant enough to base treatment decisions?
- How does my patient's unique physiology and comorbidities affect choice of medications?
- Do non-physiological considerations (cost, availability, adverse effects) play a role?

References:

1. American Diabetes Association. Summary of Revisions: *Standards of Medical Care in Diabetes-2021*. *Diabetes Care*. 2021 Jan;44(Suppl 1):S4-S6. doi: 10.2337/dc21-Srev. PMID: 33298411.
2. Hinnen D. Glucagon-Like Peptide 1 Receptor Agonists for Type 2 Diabetes. *Diabetes Spectr*. 2017 Aug;30(3):202-210. doi: 10.2337/ds16-0026. PMID: 28848315; PMCID: PMC5556578.
3. Aroda VR, Rosenstock J, Terauchi Y, et al. PIONEER 1: Randomized Clinical Trial of the Efficacy and Safety of Oral Semaglutide Monotherapy in Comparison With Placebo in Patients With Type 2 Diabetes. *Diabetes Care*. 2019;42(9):1724-1732. doi:10.2337/dc19-0749
4. Marso SP, Daniels GH, Brown-Frandsen K, et al. Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes. *N Engl J Med*. 2016;375(4):311-322. doi:10.1056/NEJMoa1603827
5. Gerstein HC, Colhoun HM, Dagenais GR, et al. Dulaglutide and cardiovascular outcomes in type 2 diabetes (REWIND): a double-blind, randomised placebo-controlled trial. *Lancet*. 2019;394(10193):121-130. doi:10.1016/S0140-6736(19)31149-3
6. Marso SP, Bain SC, Consoli A, et al. Semaglutide and Cardiovascular Outcomes in Patients with Type 2 Diabetes. *N Engl J Med*. 2016;375(19):1834-1844. doi:10.1056/NEJMoa1607141
7. Gallwitz B. Clinical Use of DPP-4 Inhibitors. *Front Endocrinol (Lausanne)*. 2019 Jun 19;10:389. doi: 10.3389/fendo.2019.00389. PMID: 31275246; PMCID: PMC6593043.
8. Matthews DR, Paldanius PM, Proot P, et al. Glycaemic durability of an early combination therapy with vildagliptin and metformin versus sequential metformin monotherapy in newly diagnosed type 2 diabetes (VERIFY): a 5-year, multicentre, randomised, double-blind trial. *Lancet*. 2019;394(10208):1519-1529. doi:10.1016/S0140-6736(19)32131-2
9. Lee SH, Gantz I, Round E, et al. A randomized, placebo-controlled clinical trial evaluating the safety and efficacy of the once-weekly DPP-4 inhibitor omarigliptin in patients with type 2 diabetes mellitus inadequately controlled by glimepiride and metformin. *BMC Endocr Disord*. 2017;17(1):70. Published 2017 Nov 6. doi:10.1186/s12902-017-0219-x
10. Joy TR, McKenzie CA, Tirona RG, et al. Sitagliptin in patients with non-alcoholic steatohepatitis: A randomized, placebo-controlled trial. *World J Gastroenterol*. 2017;23(1):141-150. doi:10.3748/wjg.v23.i1.141

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

11. Rosenstock J, Perkovic V, Johansen OE, et al. Effect of Linagliptin vs Placebo on Major Cardiovascular Events in Adults With Type 2 Diabetes and High Cardiovascular and Renal Risk: The CARMELINA Randomized Clinical Trial. *JAMA*. 2019;321(1):69-79. doi:10.1001/jama.2018.18269
12. Scirica BM, Bhatt DL, Braunwald E, et al. Saxagliptin and cardiovascular outcomes in patients with type 2 diabetes mellitus. *N Engl J Med*. 2013;369(14):1317-1326. doi:10.1056/NEJMoa1307684
13. Tentolouris A, Vlachakis P, Tzeravini E, Eleftheriadou I, Tentolouris N. SGLT2 Inhibitors: A Review of Their Antidiabetic and Cardioprotective Effects. *Int J Environ Res Public Health*. 2019 Aug 17;16(16):2965. doi: 10.3390/ijerph16162965. PMID: 31426529; PMCID: PMC6720282.
14. Stenlöf K, Cefalu WT, Kim KA, et al. Efficacy and safety of canagliflozin monotherapy in subjects with type 2 diabetes mellitus inadequately controlled with diet and exercise. *Diabetes Obes Metab*. 2013;15(4):372-382. doi:10.1111/dom.12054
15. Häring HU, Merker L, Seewaldt-Becker E, et al. Empagliflozin as add-on to metformin in patients with type 2 diabetes: a 24-week, randomized, double-blind, placebo-controlled trial. *Diabetes Care*. 2014;37(6):1650-1659. doi:10.2337/dc13-2105
16. Bhatt DL, Szarek M, Steg PG, et al. Sotagliflozin in Patients with Diabetes and Recent Worsening Heart Failure. *N Engl J Med*. 2021;384(2):117-128. doi:10.1056/NEJMoa2030183
17. Neal B, Perkovic V, Mahaffey KW, et al. Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes. *N Engl J Med*. 2017;377(7):644-657. doi:10.1056/NEJMoa1611925
18. Perkovic V, Jardine MJ, Neal B, et al. Canagliflozin and Renal Outcomes in Type 2 Diabetes and Nephropathy. *N Engl J Med*. 2019;380(24):2295-2306. doi:10.1056/NEJMoa1811744
19. Nanjan MJ, Mohammed M, Prashantha Kumar BR, Chandrasekar MJN. Thiazolidinediones as antidiabetic agents: A critical review. *Bioorg Chem*. 2018 Apr;77:548-567. doi: 10.1016/j.bioorg.2018.02.009. Epub 2018 Feb 12. PMID: 29475164.
20. Aronoff S, Rosenblatt S, Braithwaite S, Egan JW, Mathisen AL, Schneider RL. Pioglitazone hydrochloride monotherapy improves glycemic control in the treatment of patients with type 2 diabetes: a 6-month randomized placebo-controlled dose-response study. The Pioglitazone 001 Study Group. *Diabetes Care*. 2000 Nov;23(11):1605-11. doi: 10.2337/diacare.23.11.1605. PMID: 11092281.
21. Dormandy JA, Charbonnel B, Eckland DJ, et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events): a randomised controlled trial. *Lancet*. 2005;366(9493):1279-1289. doi:10.1016/S0140-6736(05)67528-9
22. Lincoff AM, Wolski K, Nicholls SJ, Nissen SE. Pioglitazone and risk of cardiovascular events in patients with type 2 diabetes mellitus: a meta-analysis of randomized trials. *JAMA*. 2007 Sep 12;298(10):1180-8. doi: 10.1001/jama.298.10.1180. PMID: 17848652.
23. Cusi K, Orsak B, Bril F, et al. Long-Term Pioglitazone Treatment for Patients With Nonalcoholic Steatohepatitis and Prediabetes or Type 2 Diabetes Mellitus: A Randomized Trial. *Ann Intern Med*. 2016;165(5):305-315. doi:10.7326/M15-1774

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

24. Rodbard HW, Rosenstock J, Canani LH, et al. Oral Semaglutide Versus Empagliflozin in Patients With Type 2 Diabetes Uncontrolled on Metformin: The PIONEER 2 Trial. *Diabetes Care*. 2019;42(12):2272-2281. doi:10.2337/dc19-0883
25. Pratley R, Amod A, Hoff ST, et al. Oral semaglutide versus subcutaneous liraglutide and placebo in type 2 diabetes (PIONEER 4): a randomised, double-blind, phase 3a trial [published correction appears in *Lancet*. 2019 Jul 6;394(10192):e1]. *Lancet*. 2019;394(10192):39-50. doi:10.1016/S0140-6736(19)31271-1
26. Weinstock RS, Guerci B, Umpierrez G, Nauck MA, Skrivaneck Z, Milicevic Z. Safety and efficacy of once-weekly dulaglutide versus sitagliptin after 2 years in metformin-treated patients with type 2 diabetes (AWARD-5): a randomized, phase III study. *Diabetes Obes Metab*. 2015;17(9):849-858. doi:10.1111/dom.12479

Submitted 5/21/2021

VI.2 Utilizing Evidence-Based Medicine for Rare Diseases (Kyla Rodgers, GSM4)

The Public Health Burden of Rare Disease

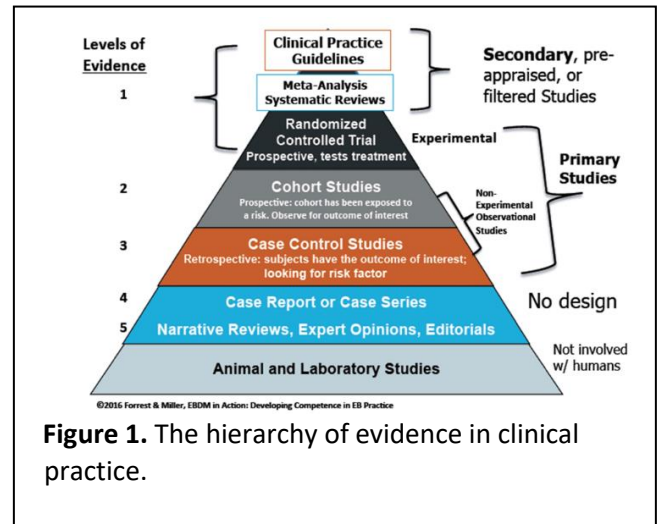
There is no singular, universally agreed upon definition of what constitutes a “rare disease”. In the United States, a disease is considered “rare” if it affects fewer than 200,000 people in the country, a definition that was set forth by the Orphan Drug Act in 1983.² Europe and Australia define a rare disease as one that affects 1 in 2000 people, while Japan and South Korea consider rare diseases those that affect fewer than 50,000 and 20,000 people in their respective countries.¹⁻³ Despite the inconsistent definition, there are over 7,000 diseases worldwide that have been classified as rare, most of which are genetic diseases and thus often affect pediatric populations¹⁻³. Taken all together, it is estimated that rare diseases affect 6-10% of the global population¹⁻³. Less than 10% of rare diseases have an available therapy, and it is estimated that only ~22% of them have ever been studied in drug trials⁵. Taken together, rare diseases are a major public health problem, particularly when considering that patients often must be seen by multiple physicians (usually over a span of years) before getting an initial diagnosis¹. Management of rare diseases can also require a tremendous number of resources. Despite the clear burden that these diseases have on public health and healthcare systems in aggregate, there is very little research activity dedicated toward discovering novel therapies to ameliorate these diseases.

Challenges in Rare Disease Research

When we learn about evidence-based medicine, we learn that there is a hierarchy of evidence (Figure 1), and that we should consider where on this hierarchy that novel clinical information falls when determining its relevance and strength. However, when it comes to rare diseases, this paradigm can become challenging. By their very nature, rare diseases affect a relatively small number of people, and recruiting these individuals to clinical trials is inherently challenging. One study in 2014 demonstrated that between January 1, 2010, and December 31, 2012 there were 659 rare disease trials, representing 70,305 enrolled patients; of these, 30.2% were discontinued, and lack of patient accrual

was the most frequently cited reason for trial discontinuation (32.1%)⁵. Of completed trials, the median number of enrolled participants was 61, with 74.5% of completed trials consisting of fewer than 100 total patients. It was also shown that trial results frequently had long delays to public disbursement, with a median time to publication of 26 months; 66.5% of trials were unpublished at 2 years, and 31.5% unpublished at 4 years after trial completion. This study demonstrates one of the unique challenges that rare diseases face when it comes to conducting research that will provide strong evidence for clinical practice recommendations: it is difficult to conduct a sufficiently powered study on a rare disease utilizing traditional research methods.

In addition to patient sparsity, there are a slew of other challenges in rare disease research. Lack of knowledge by even so-called experts or other highly trained clinicians and researchers is a major barrier to progress. Often, clinical expertise is informed by only a relatively small cohort of patients under any clinician's care. Given known major delays in diagnosis for many patients, there is tremendous gap in knowledge of the natural clinical history of rare disease. Additionally, there can be heterogeneity in clinical phenotype that is not well appreciated, again due to rarity of cases and delays in diagnoses. In turn these gaps in knowledge make it difficult to create validated disease severity scores or determine other metrics necessary for primary and secondary endpoints to measure the impact of experimental therapies in clinical trials. Owing to all these challenges, whether principles of evidence-based medicine even have a role in the approach to rare diseases has been called into question⁶.

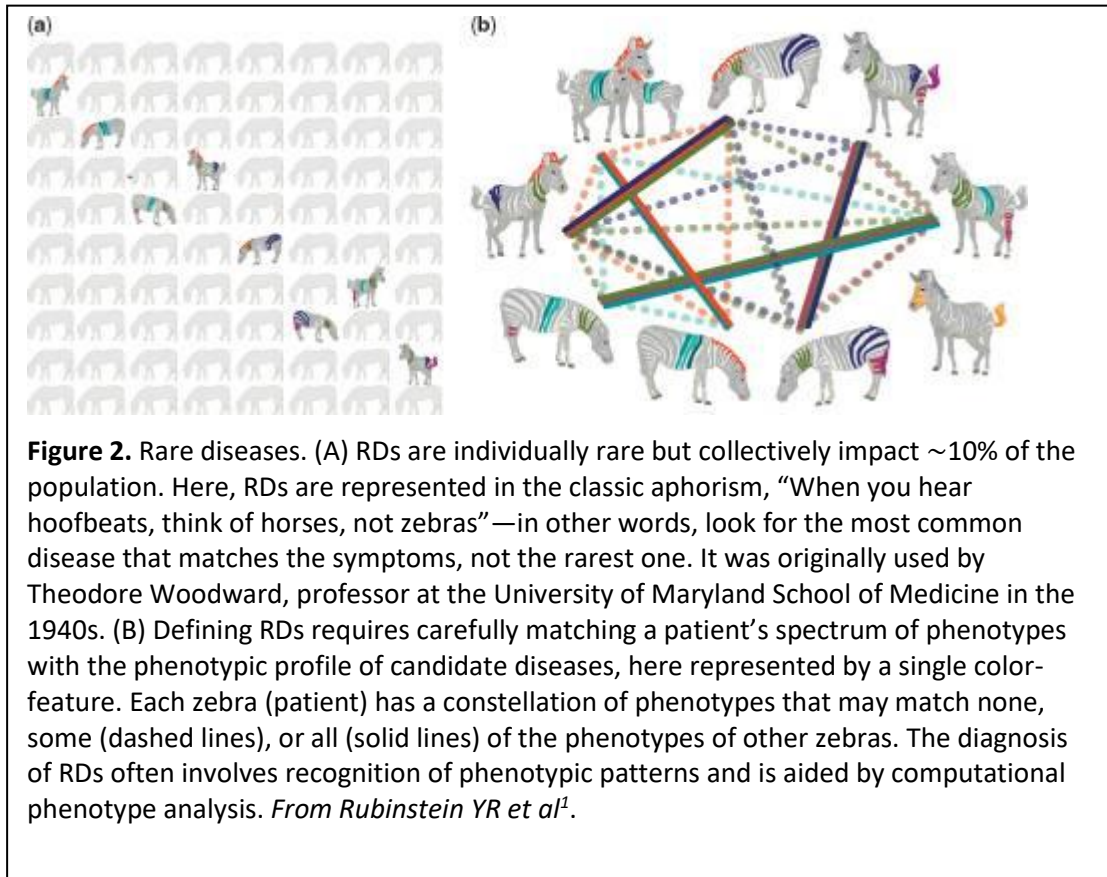


Adapting Biomedical Research to the Rare Disease Landscape

For the practitioner who endeavors to apply evidence-based medicine to the treatment of rare disease, there are several possible approaches. One is to simply utilize the available EBM hierarchy of evidence framework, with the understanding that studies may be limited and that they may have to draw conclusions from sparse data (case reports, case series), and that any RCTs they find will be considerably smaller than those of rare diseases, and thus prone to the kinds of statistical errors that plague trials with a small n (i.e. increased risk of type 2 error). For the practitioner who is simply looking to the literature to support their clinical decision making, this may be the only option. However, for the practitioner who is also looking to engage in research, there are other options that involve novel ways of approaching biomedical research, utilizing unique clinical trial designs for example.

A Novel Approach to Biomedical Research

The biomedical research paradigm for decades has been largely a “top-down” driven approach. In this model, a physician and/or scientist who plays the role of expert identifies a major need or gap in medical knowledge regarding a particular patient group. This expert compiles the evidence, writes a research proposal for a clinical trial, and procures funding via a government, academic, or industrial source. Patients are recruited to the trial and serve as data points for the research; typically, these patients are drawn from a community surrounding a particular academic center that is serving as a site for a clinical trial. This process is quite passive from the patient’s perspective, and it works well for very common diseases. However, this traditional model fails when it comes to rare diseases for several important reasons. One is that it relies on the physician/scientist to be an expert on the disease being studied; as we have already established, this is not often the case when it comes to rare diseases. The second is that it relies on being able to recruit patients to the study, which relies on having patients near a participating clinical site; given the low density of patients with the specific rare disease under study, this is another condition that is difficult to fulfil.



So how do we solve these problems? One possibility that is quickly gaining traction is to adopt a “bottom-up” approach^{3,7}. In this model, patients and their families form a network and/or foundation dedicated to their rare disease of interest, discuss their experiences to build consensus on disease phenotype (Figure 2) as well as identify shared major challenges, then identify physicians/scientists who are best positioned to conduct the research. This paradigm is significantly more patient-centered than the traditional biomedical research model, as it recognizes that patients (particularly those with rare disease) are experts on the physical (not to mention psychological and emotional) manifestations of their disease, as well as what kinds of goals they hope to meet with novel therapeutics. Studies have shown that patients with rare diseases are keen to share their health data with researchers, with 90-97% of respondents endorsing a willingness to share health data to help researchers better understand mechanisms of disease, develop new treatments, improve diagnosis, receive additional specialist advice on care, and even to improve research and care of diseases other than their own⁸.



In the internet age, building a patient network has become significantly easier. Patient-initiated support groups can be found through several major social media platforms, including Facebook Groups and Reddit “subs”, as well as via Twitter (via the use of hashtags and dedicated accounts). These groups have been successfully utilized to recruit patients to help build disease-specific registries, facilitate discussions on community needs, etc. From these spaces, dedicated disease organizations (such as the Castleman Disease Collaborative Network³) have been successfully launched.

This all sounds ideal, but how does this work, practically speaking? The power of the patient-driven disease network to recruit participants to a clinical trial was clearly demonstrated in the case of alkaptonuria (AKU) and a drug called nitisinone. This drug was particularly promising for the treatment of AKU, as it blocks the enzyme 4-Hydroxyphenylpyruvate dioxygenase (HPPD), leading to a 95% reduction in the level of homogentisic acid, the substance that builds up in connective tissues in patients with AKU. An RCT was run by the NIH in 2008 in which the drug failed to meet its clinical endpoint; one possible reason cited was that only 40 patients were included in the trial, half of which received the drug, as well as a relatively narrow clinical endpoint of hip mobility^{3,9}. However, the AKU Society subsequently elected to fund a four-year study that was designed by a team of patients and physicians, who worked together to build a disease severity index to track disease progression; this trial was able to recruit 3.45 times more patients (138 total) than the investigator-initiated NIH trial^{3,10}. The results of this trial were positive, and nitisinone was subsequently approved for use in AKU in Europe. The example of nitisinone is a powerful case study in how patient-centered groups can shape the international research agenda and successfully develop novel therapies for rare disease, while still adhering to the high standards of EBM.

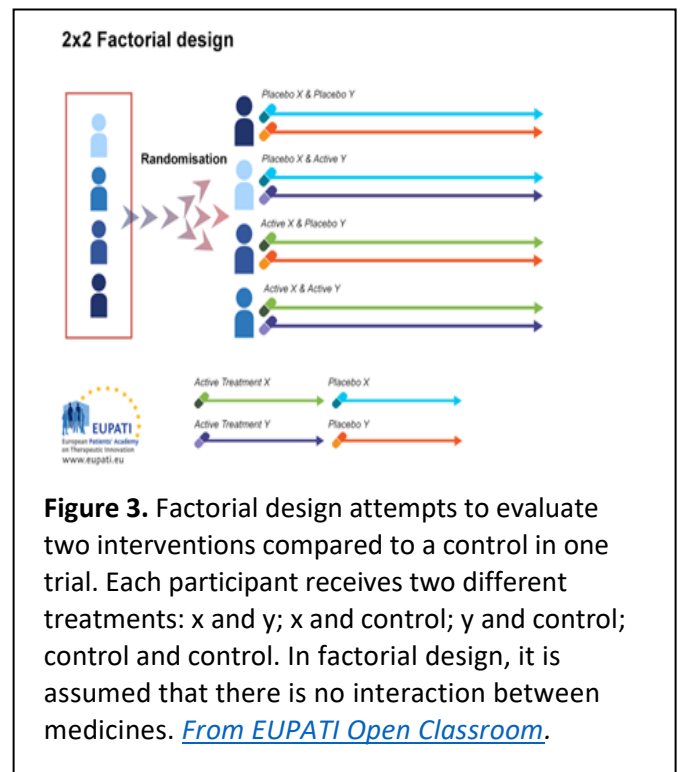


Figure 3. Factorial design attempts to evaluate two interventions compared to a control in one trial. Each participant receives two different treatments: x and y; x and control; y and control; control and control. In factorial design, it is assumed that there is no interaction between medicines. [From EUPATI Open Classroom.](#)

Innovative Trial Designs

In addition to utilizing patient registries and robust social/support networks to increase patient recruitment to clinical trials, as well as very careful consideration of appropriate primary endpoints for measuring efficacy, some have suggested utilizing novel clinical trial designs^{11,12}. As Gagne et al pointed out, there are essentially two approaches to modifying clinical trials for a rare disease population: either minimizing the size of the trial in a way that allows researchers to extract clinically meaningful data from as few patients as possible, or maximizing the n of the trial as much as possible in order to minimize biases and random errors, which are more likely to have an outsized effect on the interpretation of the data in a small population^{11,12}. The following is a summary of the most salient points regarding the use of novel trial designs and research methods in rare diseases, from their excellent study¹¹.

Minimizing Trial Size

- **Factorial design:** *These trials are designed to answer multiple questions within a study population, while minimizing the number of participants required to answer these multiple questions (Figure 3)¹¹. Factorial design has been reviewed in depth in a separate chapter in this guide by Yi Zhang.*
- **Adaptive randomization:** *Adaptive randomization is a design that allows for changes in patient assignment based on interim analysis of trial data. One type is called covariate-adaptive randomization, which allows investigators to change patient assignment in order to ensure more balanced baseline characteristics between treatment arms. Another type is called response-adaptive randomization (Figure 4); this type of trial allows investigators to reassign patients between treatment arms based on safety and efficacy data. In the example shown in Figure 4, a theoretical drug is tested at three doses; in the interim analysis, the highest dose is shown to be associated with higher risk and more serious side effects. This allows investigators to preferentially assign newly recruited patients (or current participants) to the medium and low dose arms, which show both better safety and higher tolerability than the high dose. One danger of this type of design is that, without prior knowledge of the therapy and/or patient population, it can be quite difficult to determine when the first interim analysis should take place. An inappropriately timed interim analysis could result in either a missed opportunity to assign patients to a more promising therapy and/or unnecessary harm.*

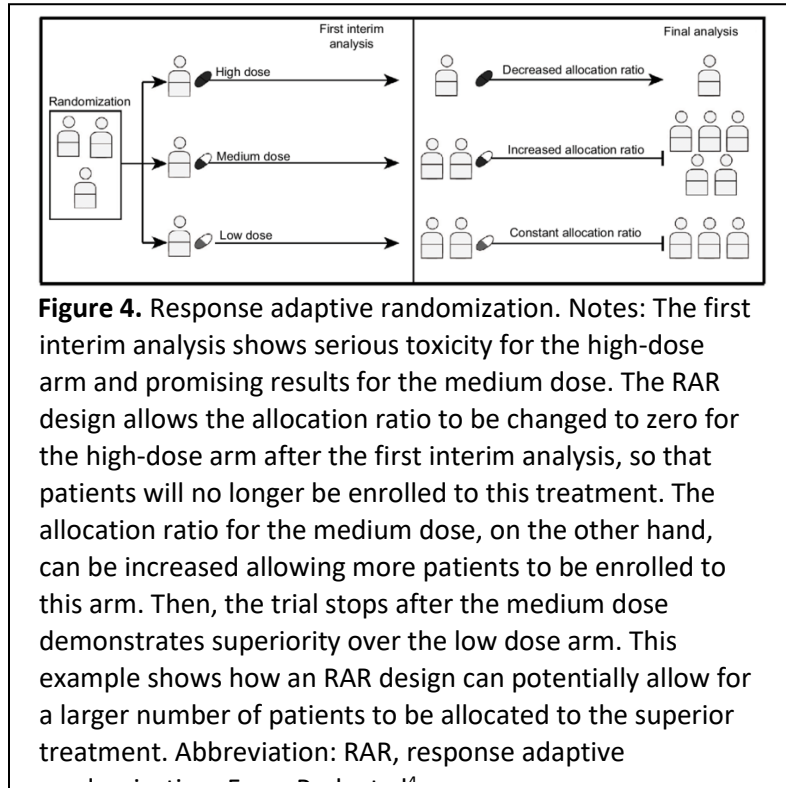


Figure 4. Response adaptive randomization. Notes: The first interim analysis shows serious toxicity for the high-dose arm and promising results for the medium dose. The RAR design allows the allocation ratio to be changed to zero for the high-dose arm after the first interim analysis, so that patients will no longer be enrolled to this treatment. The allocation ratio for the medium dose, on the other hand, can be increased allowing more patients to be enrolled to this arm. Then, the trial stops after the medium dose demonstrates superiority over the low dose arm. This example shows how an RAR design can potentially allow for a larger number of patients to be allocated to the superior treatment. Abbreviation: RAR, response adaptive

- **Sequential trials:** In sequential trials (Figure 5), there are pre-planned interim analyses of safety and efficacy data. This design allows trials to be stopped early if there is evidence of either significant harm or benefit; if either of these are found, then clinical equipoise no longer exists, and ethically speaking the RCT should be halted. This design does not preferentially shuttle patients to a different arm in a semi-random manner the way that response-adaptive randomization does, but *if* a trial is stopped early, then it utilizes fewer patients. Of note, this design only minimizes patient number if it meets conditions for early termination.

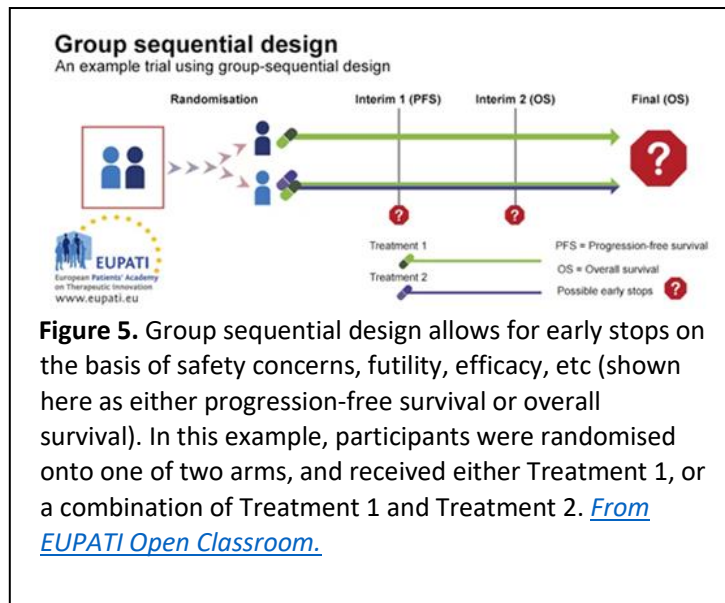


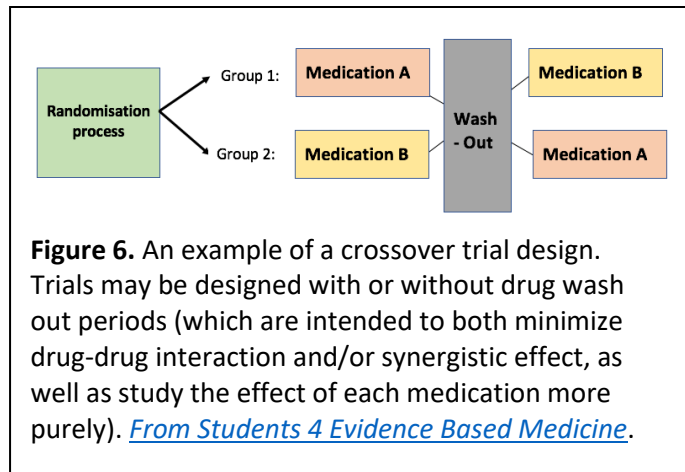
Figure 5. Group sequential design allows for early stops on the basis of safety concerns, futility, efficacy, etc (shown here as either progression-free survival or overall survival). In this example, participants were randomised onto one of two arms, and received either Treatment 1, or a combination of Treatment 1 and Treatment 2. [From EUPATI Open Classroom.](#)

- **Enhancing statistical power:** As discussed above, the choice of primary outcome for a clinical trial is very important. It is critical to choose an endpoint that is both easily measurable and clinically significant to ensure that you are measuring a real change in disease activity that will have meaningful clinical change for the patient, and that the data are reliable. Some ways to enhance statistical power includes use of a:
 - **continuous outcome variable** (a variable that can have any continuous variable, such as height or BMI)
 - **surrogate market** (i.e., biomarker, though these are difficult to validate in rare disease populations)
 - **composite endpoint** (multiple endpoints are combined into one single score/endpoint)
 - **repeated measure outcomes** (the same measurement made serially in time, or under various experimental conditions)
- **Re-considering standard statistical approaches:** An important consideration for not only rare diseases, but common ones as well, is whether "statistical significance" is an appropriate marker for clinical significance. Conventionally most clinical trials and basic science research utilize an α of 0.05 to determine whether to reject the null hypothesis or not. However, there is a movement to reconsider whether reaching a p-value of <0.05 always indicates a clinically meaningful response, and, vice-versa, whether failing to reach a p-value of <0.05 always indicates a lack of clinically meaningful response¹³. With that in mind, there are several possible approaches to statistical analysis in rare disease clinical trials:

- Increase α
- Conduct an underpowered study with the intent to incorporate results into a prospective meta-analysis
- Incorporate the results into a Bayesian framework (Bayes theorem and its strengths and pitfalls are thoroughly reviewed elsewhere in this EBM Guide by Malachy Sullivan, Alex Briand, and Stephen Conn; please see these chapters for further detail)

Maximizing Trial Size

- **Crossover trials:** In these trials (Figure 6), participants spend some length of time on either the treatment under study or placebo (or other comparator), then switch to the other arm of the study for another length of time. This allows patients to serve as their own control, minimizes concerns regarding balancing baseline characteristics of patients between arms, and allows investigators to get on-treatment data from *all* participants, rather than half of



This design can be considered either a way of minimizing trial size (you can get the same amount of data from half of the number of patients as a conventional RCT), or you can consider it a way of maximizing trial size, as you are also effectively doubling the amount of on-treatment data that one would get from the same number of patients in a conventional RCT. There are several variations of the crossover trial, including *n-of-1* trials (which can be planned as part of a prospective meta-analysis), an alternating design (in which patients crossover between treatments multiple times over a period of time, such that treatments are alternating), and other less common designs, including Latin square, stepped wedge, and randomized withdrawal designs, which are beyond the scope of this review.

Summary

Rare diseases represent a unique challenge for the EBM framework; patients are few and far between, the knowledge base of most clinicians/researchers is scarce due to lack of exposure to these diseases, and study design is a unique challenge. However, the rise of patient registries, including government-initiated organizations such as the NIH Undiagnosed Diseases Network, NGOs such as National Organization of Rare Disease (NORD), and patient-initiated organizations such as the Castleman Disease Collaborative Network, are critical for framing problems, forming meaningful research questions, and recruiting patients for clinical trials. Combining innovative clinical trial design and research methods with these powerful networks is the key to generating strong, evidence-based advances in rare disease research.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

References/Footnotes:

1. Rubinstein YR, Robinson PN, Gahl WA, et al. The case for open science: rare diseases. *JAMIA Open* 2020;3:472-86.
2. Mitani AA, Haneuse S. Small Data Challenges of Studying Rare Diseases. *JAMA Netw Open* 2020;3:e201965.
3. Ainsworth C. Rare diseases band together toward change in research. *Nature Medicine* 2020;26:1496-9.
4. Park JJ, Thorlund K, Mills EJ. Critical concepts in adaptive clinical trials. *Clin Epidemiol* 2018;10:343-51.
5. Rees CA, Pica N, Monuteaux MC, Bourgeois FT. Noncompletion and nonpublication of trials studying rare diseases: A cross-sectional analysis. *PLoS Med* 2019;16:e1002966.
6. Kruer MC, Steiner RD. The role of evidence-based medicine and clinical trials in rare genetic disorders. *Clin Genet* 2008;74:197-207.
7. Rath A, Salamon V, Peixoto S, et al. A systematic literature review of evidence-based clinical practice for rare diseases: what are the perceived and real barriers for improving the evidence and how can they be overcome? *Trials* 2017;18:556.
8. Courbier S, Dimond R, Bros-Facer V. Share and protect our health data: an evidence based approach to rare disease patients' perspectives on data sharing and data protection - quantitative survey and recommendations. *Orphanet J Rare Dis* 2019;14:175.
9. Introne WJ, Perry MB, Troendle J, et al. A 3-year randomized therapeutic trial of nitisinone in alkaptonuria. *Mol Genet Metab* 2011;103:307-14.
10. Ranganath LR, Psarelli EE, Arnoux JB, et al. Efficacy and safety of once-daily nitisinone for patients with alkaptonuria (SONIA 2): an international, multicentre, open-label, randomised controlled trial. *Lancet Diabetes Endocrinol* 2020;8:762-72.
11. Gagne JJ, Thompson L, O'Keefe K, Kesselheim AS. Innovative research methods for studying treatments for rare diseases: methodological review. *BMJ : British Medical Journal* 2014;349:g6802.
12. Behera M, Kumar A, Soares HP, Sokol L, Djulbegovic B. Evidence-based medicine for rare diseases: implications for data interpretation and clinical trial design. *Cancer Control* 2007;14:160-6.
13. Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: Clinical versus statistical significance. *Perspect Clin Res* 2015;6:169-70.

Submitted 1/2/2022

VI.3 Understanding Falls in the Elderly Utilizing EBM (Xingyi Li, GSM4)

Falls are a common problem for older adults that significantly affect mortality, morbidity and quality of life. Studies have shown that as many as one third of older adults experience at least one fall every year. The subsequent fear of falling and physical deconditioning, in turn, further increase the risk for future falls [1]. Many patients have to go through institutionalization and a long-term rehabilitation process before, if they could ever, returning to their previous level of independence. Falls are also associated

with high depression rate and functional decline [2]. Additionally, a large proportion of healthcare spending every year is associated with older adult falls. According to a 2018 study, every year \$50 billion is spent on non-fatal falls of adults over age 65 and \$754 million is spent on fatal falls [3]. As a result, fall prevention plays an important role in geriatric care, both for patients and for the healthcare system as a whole.

Risk factors for falls in older adults are often multifactorial. According to a systematic review, the major risk factors include impaired gait, polypharmacy and history of previous falls. Comorbidities such as visual impairment and cognitive impairment cannot be overlooked when assessing a patient's risk of sustaining a fall [4]. Sometimes falls can also be a manifestation of another illness. For example, around 20% of all cardiovascular syncope in patients over 70 years old presents as falls [5].

The American Geriatrics Society published a diagnostic algorithm for post-fall assessment for older adults that incorporated the multifactorial nature of falls (figure 1) [6]. The National Institute on Aging (NIA) also made a few recommendations for the elderly to reduce the risk of falling [7]. This chapter uses the skills taught in the Evidence Based Medicine Elective to evaluate and elaborate on some of these recommendations. Asking questions leads one to pursue best evidence, followed by integrating an approach based on such evidence.

Is staying physically active an effective method of preventing falls in older adults?

Exercise is an intervention that can effectively prevent falls in community-dwelling older people. A British meta-analysis of 88 randomized control trials with 19,478 participants from January 2010 to January 2016 showed that exercise group had a 21% reduction in fall rate (pooled rate ratio 0.79, 95% CI 0.73 to 0.85, $p < 0.001$, I² 47%, 69 comparisons). Among all different types of exercise program designs, those involving more than 3 hours/week of exercise and balance training had the most significant effect. Different comorbidities, however, were associated with different results. The beneficial effects of exercise were demonstrated in people with Parkinson's disease (pooled rate ratio 0.47, 95% CI 0.30 to 0.73, $p = 0.001$, I² 65%, 6 comparisons) or cognitive impairment (pooled rate ratio 0.55, 95% CI 0.37 to 0.83, $p = 0.004$, I² 21%, 3 comparisons) but not people who recently survived a stroke or were discharged from hospital [8].

It is worth pointing out that a reduction in the number of falls does not necessarily mean there is a reduction in the number of fallers. Another meta-analysis that included 25 randomized control trials with people with Parkinson's disease showed that exercise-enhancing balance and gait performance led to a reduction in number of falls over a short and long period of time, but there was no evidence of decreased number of fallers [9].

Exercise is an umbrella term and encompasses a variety of different trainings. For the purpose of fall prevention, more than one study has shown the following exercises were the most helpful [8,10]:

- Resistance training at least two to three times per week
- Endurance training with increasing intensity as tolerated
- Balance training with several exercise stimuli
- At least 3 hours of training each week
- Continued training is required, otherwise benefits will be lost

Are exercise programs implemented with telehealth effective in fall prevention?

With technology advancement, telehealth is becoming a more and more important format of care delivery, especially in the era of the COVID pandemic. Exercise programs are usually long-term interventions carried out in the community setting. Therefore, telehealth provides a unique opportunity to relieve the older patients from the burden of transportation, as well as to engage in exercise in their home environment. A feasibility study that looked into the application of video-enhanced care management in older veterans found that the complexity of video-enhanced care is generally accepted among participants and it is feasible to use video conference as a healthcare delivery method [11]. The benefits of telehealth fall prevention programs are supported by RCTs. A two-year RCT of 503 participants showed that a home-based e-health balance exercise program significantly reduced the rate of falls over the two years (incidence rate ratio 0.84, 95% confidence interval 0.72 to 0.98, P=0.027) [12].

In addition to fall prevention, telehealth interventions were proven to be useful for other outcomes. For example, a RTC of 115 participants has found that an integrated telehealth service can effectively lower ED usage for older adults with CHF or COPD [13]. It was also shown that telehealth service as a support for self-monitoring may reduce the number of hospitalizations among older adults with multiple comorbidities [14].

Is addressing vision problems an effective method of preventing falls in older adults?

Studies have shown that impaired vision (e.g. impaired depth perception, low color sensitivity, low contrast visual acuity are the strongest risk factors) is an important and independent risk factor for falls among older adults [15]. However, the evidence for the benefits of visual correction is mixed. One randomized control trial of 276 participants did not show a significant reduction in falls [16]. Moreover, a trial involving 616 people over age 70 showed a higher fall rate in people receiving vision intervention (65% fell at least once; 758 falls in total) compared to the control group (50% fell at least once; 516 falls in total), possibly due to need for adjustment and recovery post treatment [17]. Another RCT with 306 women aged over 70 showed that cataract surgery reduced the rate of falling by 35% (rate ratio 0.66, 95% CI 0.45 to 0.96, p = 0.03), but there was no significant reduction in the number of fallers (49% fell at least once in the intervention group and 45% fell at least once in the control group) [18].

The challenge of practicing EBM for this clinical problem is a lack of evidence. There are not enough studies on visual correction as a fall prevention intervention alone. Most studies implemented visual screening/correction as a part of a multi-factorial intervention and the results are generally mixed. Based on these data, addressing vision problems for older adults may be a beneficial intervention for fall prevention, but it alone is not enough to have significant clinical effect.

Is addressing polypharmacy problem an effective method of preventing falls in older adults?

Many medications are associated with higher risk for falls in older adults, especially psychotropic medications which are associated with an as high as a 47% increase in fall rate among older adults living in the community [4]. Polypharmacy of cardiovascular drugs is also shown to increase the risk of falling [19]. The exact mechanism of how polypharmacy increases risk for falls is not clearly known. One trial found that using more than two fall-risk-increasing drugs, rather than polypharmacy, increased the risk for falls [20], which may indicate that the exact mechanism of a specific drug is more important than the number of drugs used in clinical practice.

A RTC with 93 older adults showed that withdrawal from psychotropic medication alone was an effective fall prevention intervention over 44 weeks (relative risk 0.34, 95% CI, 0.16-0.74) [21]. Some multi-factorial programs included medication optimization as a part of the intervention and had good results [22-24]. However, due to the nature of multi-factorial intervention, whether addressing the medication list was associated with any significant benefit cannot be assessed with these studies.

Compared to interventions involving visual correction, there are more data on medication optimization, including studies using medication withdrawal as a sole intervention and studies that address polypharmacy as a part of a multi-factorial intervention program. There are also more studies approved and will be conducted in the near future on a protocol of addressing polypharmacy [25]. Hopefully clinicians will have more evidence to guide practice.

Is having good footwear an effective method of preventing falls in older adults?

An evaluation of risk factors for falls stated that foot problems have an added odds ratio for falls of 1.8. Foot problems include moderate or severe bunions, toe deformities, ulcers, or deformed nails [26]. Inappropriate footwear is one of the common causes for foot problems [27]. Studies have found that 75% older adults admitted for hip fracture were wearing hazardous footwear such as sandals and slippers when they fell [28].

Are appropriate, or fall-preventing footwear an effective method of preventing falls? A RCT of 44 community-dwelling people over age 65 showed that there was no significant reduction in the fall incidence of using bilateral custom-made ankle-foot orthoses [29]. Another literature review study on

non-slip socks found no evidence supporting the benefits of its usage in hospitals [30]. Overall, it is inconclusive whether appropriate footwear alone is effective in reducing fall risks among older adults.

There is not adequate evidence directly on the topic of whether appropriate footwear alone would reduce the risk of falls in older adults. Some multi-factorial fall prevention interventions included providing new footwear as a part of the program [31,32]. There are also trials evaluating podiatry interventions which included foot and ankle exercises, foot orthoses and new footwear for participants. It was found that comprehensive podiatry care effectively reduces the incidence of falls among older adults [33], but appropriate footwear, just like in the multi-factorial intervention studies, is only one part of the intervention program.

Is limiting alcohol intake an effective method of preventing falls in older adults?

Alcohol can affect one's judgment, coordination and reaction time. Long term use may lead to long-lasting change in nervous system chemistry. A literature review of 182 published articles found that the likelihood of falling is directly associated with the amount of alcohol consumed [34]. In addition, a study that reviewed over 38,000 ED records found that alcohol-related fall presentations as ED visits are more likely to result in traumatic brain injuries than those visits with no alcohol indication (34.8% vs 17%) [35].

Despite a strong consensus on the harmful nature of alcohol consumption, there is limited data on whether decreasing alcohol consumption alone is an effective intervention for fall prevention.

Is home environment modification an effective method of preventing falls in older adults?

Environmental hazards can pose additional risks to older adults. It is recommended that a fall prevention program include home environment screening and modification (hazard identification and removal, installation of handrails and grab bars, and improvement of lighting) [36].

The challenge of evaluation of this recommendation using the EBM skill is similar to the previous ones – the lack of studies on home modification as a single fall prevention intervention. However, the evidence of multi-factorial interventions is strong. A meta-analysis of 33 multi-factorial trials found significant benefit of the interventions that included home screening and modification, along with exercise programs, education, vision and medication screening [37].

Summary:

Exercise is the single most effective intervention for fall prevention among older adults. Many studies support the benefits of a structured, long-term exercise program for people with high fall risks. Endurance and balance trainings are most useful for fall prevention. Exercise programs can be conducted in the telehealth format to reduce attrition rate. In addition to exercise, research has shown

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

that a combination with other interventions such as minimization of medications and addressing vision problems may have added value for fall prevention.

One limitation of using EBM skills to review this topic is the lack of data, especially on recommendations other than exercise. Nonetheless, approaching this significant clinical problem in the context of best evidence leads one to search, find, appraise and interpret prior research. There are many studies that incorporated a multi-factorial intervention program to reduce fall incidence; yet the number of trials that look into the efficacy of a single intervention is limited. There are many possible contributing factors. First, it may not be as easy to get funding to conduct a trial that looks into, for example, vision screening as a fall prevention intervention alone. Second, conducting a trial takes a long time, ranging from a few months to years, and single-factorial trials may not be worth the time and commitment of a research team. However, even with limited data, it is still reasonable to conclude that the aforementioned recommendations are worth considering in clinical practice. There are adequate data to support poor vision, polypharmacy, inappropriate footwear, alcohol intake and hazardous home environment as risk factors for falls. The multi-factorial trials also proved these recommendations as effective interventions when combined with exercise.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

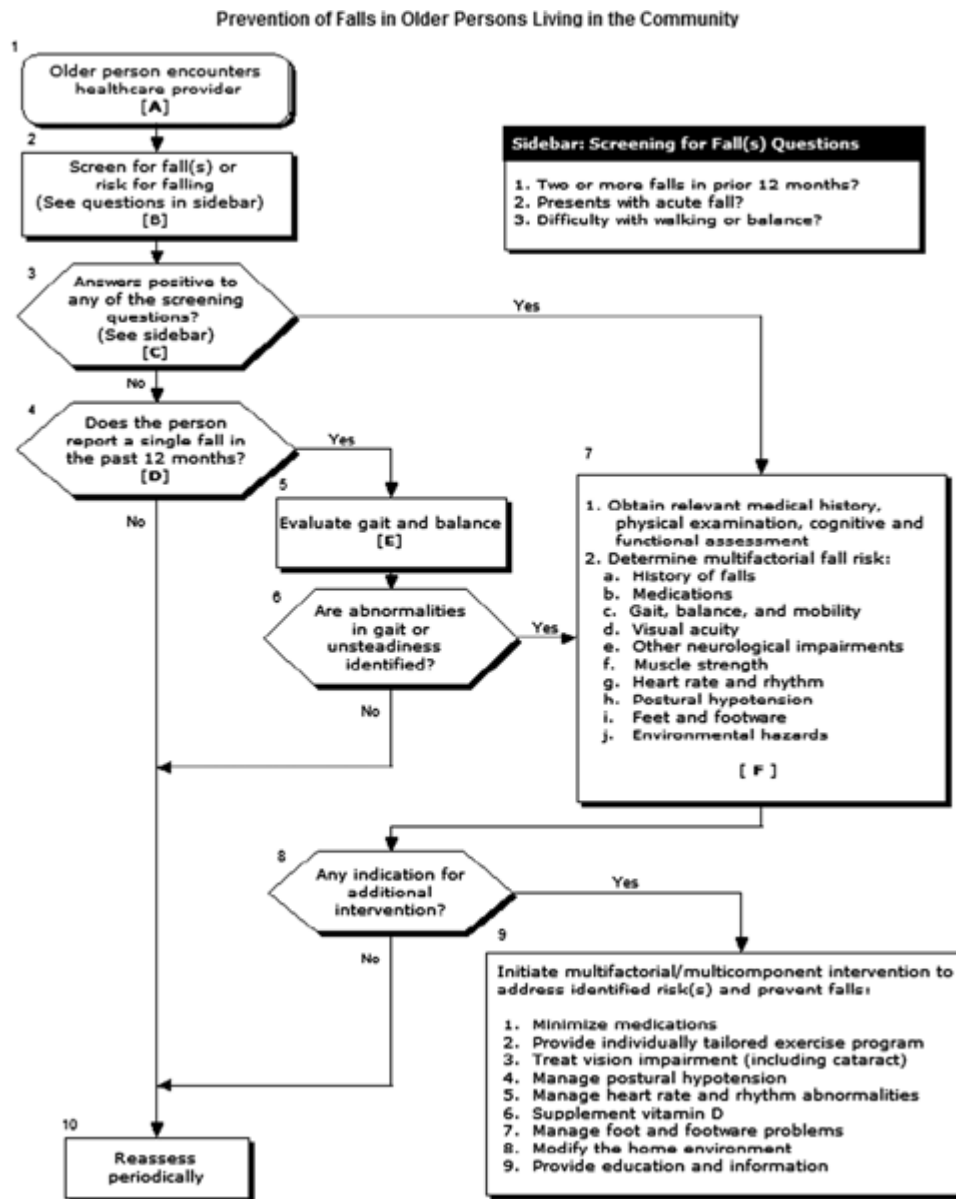


Figure 1; Credit: Summary of the Updated American Geriatrics Society/British Geriatrics Society Clinical Practice Guideline for Prevention of Falls in Older Persons. Journal of the American Geriatrics Society, 2011

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

References:

1. Lusardi MM, Fritz S, Middleton A, et al. Determining Risk of Falls in Community Dwelling Older Adults: A Systematic Review and Meta-analysis Using Posttest Probability. *J Geriatr Phys Ther.* 2017;40(1):1-36.
2. Scaf-Klomp W, Sanderman R, Ormel J, Kempen GI. Depression in older people after fall-related injuries: a prospective study. *Age Ageing.* 2003 Jan;32(1):88-94.
3. Center for Disease Control and Prevention. Cost of older Adult Falls. Center for Disease Control and Prevention. Retrieved 2022, from <https://www.cdc.gov/falls/data/fall-cost.html>
4. Ambrose AF, Paul G, Hausdorff JM. Risk factors for falls among older adults: a review of the literature. *Maturitas.* 2013 May;75(1):51-61.
5. Ungar A, Rafanelli M, Iacomelli I, et al. Fall prevention in the elderly. *Clin Cases Miner Bone Metab.* 2013;10(2):91-95.
6. Guideline for the prevention of falls in older persons. American Geriatrics Society, British Geriatrics Society, and American Academy of Orthopaedic Surgeons Panel on Falls Prevention. *J Am Geriatr Soc.* 2001 May;49(5):664-72.
7. U.S. Department of Health and Human Services. Prevent falls and fractures. National Institute on Aging. Retrieved 2022, from <https://www.nia.nih.gov/health/prevent-falls-and-fractures>
8. Sherrington C, Michaleff ZA, Fairhall N, Paul SS, Tiedemann A, Whitney J, Cumming RG, Herbert RD, Close JCT, Lord SR. Exercise to prevent falls in older adults: an updated systematic review and meta-analysis. *Br J Sports Med.* 2017 Dec;51(24):1750-1758. doi: 10.1136/bjsports-2016-096547. Epub 2016 Oct 4.
9. Shen X, Wong-Yu IS, Mak MK. Effects of Exercise on Falls, Balance, and Gait Ability in Parkinson's Disease: A Meta-analysis. *Neurorehabil Neural Repair.* 2016 Jul;30(6):512-27.
10. Cadore EL, Rodríguez-Mañas L, Sinclair A, Izquierdo M. Effects of different exercise interventions on risk of falls, gait ability, and balance in physically frail older adults: a systematic review. *Rejuvenation Res.* 2013;16(2):105-114.
11. Hastings SN, Mahanna EP, Berkowitz TSZ, Smith VA, Choate AL, Hughes JM, Pavon J, Robinson K, Hendrix C, Van Houtven C, Gentry P, Rose C, Plassman BL, Potter G, Oddone E. Video-Enhanced Care Management for Medically Complex Older Adults with Cognitive Impairment. *J Am Geriatr Soc.* 2021 Jan;69(1):77-84.
12. Delbaere K, Valenzuela T, Lord SR, Clemson L, Zijlstra GAR, Close JCT, Lung T, Woodbury A, Chow J, McInerney G, Miles L, Toson B, Briggs N, van Schooten KS. E-health StandingTall balance exercise for fall prevention in older people: results of a two year randomised controlled trial. *BMJ.* 2021 Apr 6;373:n740.
13. Gellis ZD, Kenaley BL, Ten Have T. Integrated telehealth care for chronic illness and depression in geriatric home care patients: the Integrated Telehealth Education and Activation of Mood (I-TEAM) study. *J Am Geriatr Soc.* 2014 May;62(5):889-95.
14. Lear SA, Norena M, Banner D, Whitehurst DGT, Gill S, Burns J, Kandola DK, Johnston S, Horvat D, Vincent K, Levin A, Kaan A, Van Spall HGC, Singer J. Assessment of an Interactive Digital Health-Based Self-management Program to Reduce Hospitalizations Among Patients With Multiple Chronic Diseases: A Randomized Clinical Trial. *JAMA Netw Open.* 2021 Dec 1;4(12):e2140591.
15. Lord, S.R. and Dayhew, J. (2001), Visual Risk Factors for Falls in Older People. *Journal of the American Geriatrics Society*, 49: 508-515.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

16. Gillespie LD, Gillespie WJ, Robertson MC, Lamb SE, Cumming RG, Rowe BH. Interventions for preventing falls in elderly people. *Cochrane Database Syst Rev.* 2003;(4):CD000340.
17. Cumming RG, Ivers R, Clemson L, Cullen J, Hayes MF, Tanzer M, Mitchell P. Improving vision to prevent falls in frail older people: a randomized trial. *J Am Geriatr Soc.* 2007 Feb;55(2):175-81.
18. Harwood RH, Foss AJ, Osborn F, Gregson RM, Zaman A, Masud T. Falls and health status in elderly women following first eye cataract surgery: a randomised controlled trial. *Br J Ophthalmol.* 2005 Jan;89(1):53-9.
19. Huang, AR, Mallet, L, Rochefort, CM et al. Medication-Related Falls in the Elderly. *Drugs Aging.* 2012 29, 359–376.
20. Zia A, Kamaruzzaman SB, Tan MP. The consumption of two or more fall risk-increasing drugs rather than polypharmacy is associated with falls. *Geriatr Gerontol Int.* 2017 Mar;17(3):463-470.
21. Campbell AJ, Robertson MC, Gardner MM, Norton RN, Buchner DM. Psychotropic medication withdrawal and a home-based exercise program to prevent falls: a randomized, controlled trial. *J Am Geriatr Soc.* 1999 Jul;47(7):850-3.
22. Lee HC, Chang KC, Tsauo JY, Hung JW, Huang YC, Lin SI; Fall Prevention Initiatives in Taiwan (FPIT) Investigators. Effects of a multifactorial fall prevention program on fall incidence and physical function in community-dwelling older adults with risk of falls. *Arch Phys Med Rehabil.* 2013 Apr;94(4):606-15, 615.e1.
23. Clemson L, Cumming R, Kendig H et al. The effectiveness of a communitybased program for reducing the incidence of falls in the elderly: A randomized trial. *J Am Geriatr Soc* 2004;52:1487–1494.
24. Close J, Ellis M, Hooper R et al. Prevention of falls in the elderly trial (PROFET): A randomized controlled trial. *Lancet* 1999;353:93–97.
25. Bergler U, Ailabouni NJ, Pickering JW, Hilmer SN, Mangin D, Nishtala PS, Jamieson H; Sponsor-investigator. Deprescribing to reduce polypharmacy: study protocol for a randomised controlled trial assessing deprescribing of anticholinergic and sedative drugs in a cohort of frail older people living in the community. *Trials.* 2021 Nov 3;22(1):766.
26. Tinetti ME, Speechley M, Ginter SF. Risk factors for falls among elderly persons living in the community. *N Engl J Med.* 1988 Dec 29;319(26):1701-7.
27. Shirzad K, Kiesau CD, DeOrio JK, Parekh SG. Lesser toe deformities. *J Am Acad Orthop Surg.* 2011 Aug;19(8):505-14.
28. Sherrington C, Menz HB. An evaluation of footwear worn at the time of fall-related hip fracture. *Age Ageing.* 2003 May;32(3):310-4.
29. Wang C, Goel R, Zhang Q, Lepow B, Najafi B. Daily Use of Bilateral Custom-Made Ankle-Foot Orthoses for Fall Prevention in Older Adults: A Randomized Controlled Trial. *J Am Geriatr Soc.* 2019 Aug;67(8):1656-1661.
30. Hartung B, Lalonde M. The use of non-slip socks to prevent falls among hospitalized older adults: A literature review. *Geriatr Nurs.* 2017 Sep-Oct;38(5):412-416.
31. Lightbody E, Watkins C, Leathley M, Sharma A, Lye M. Evaluation of a nurse-led falls prevention programme versus usual care: a randomized controlled trial. *Age Ageing.* 2002 May;31(3):203-10.
32. Lord SR, Tiedemann A, Chapman K et al. The effect of an individualized fall prevention program on fall risk and falls in older people: A randomized, controlled trial. *J Am Geriatr Soc* 2005;53:1296–1304
33. Cockayne S, Adamson J, Clarke A, Corbacho B, Fairhurst C, Green L, Hewitt CE, Hicks K, Kenan AM, Lamb SE, McIntosh C, Menz HB, Redmond AC, Richardson Z, Rodgers S, Vernon W, Watson J, Torgerson DJ; REFORM study. Cohort Randomised Controlled Trial of a Multifaceted Podiatry Intervention for the Prevention of Falls

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

in Older People (The REFORM Trial). *PLoS One*. 2017 Jan 20;12(1):e0168712.

34. Taylor B, Irving HM, Kanteres F, Room R, Borges G, Cherpitel C, Greenfield T, Rehm J. The more you drink, the harder you fall: a systematic review and meta-analysis of how acute alcohol consumption and injury or collision risk increase together. *Drug Alcohol Depend*. 2010 Jul 1;110(1-2):108-16.
35. Shakya I, Bergen G, Haddad YK, Kakara R, Moreland BL. Fall-related emergency department visits involving alcohol among older adults. *J Safety Res*. 2020 Sep;74:125-131.
36. Summary of the Updated American Geriatrics Society/British Geriatrics Society Clinical Practice Guideline for Prevention of Falls in Older Persons. *Journal of the American Geriatrics Society*, 2011, 59: 148-157.
37. Chase CA, Mann K, Wasek S, Arbesman M. Systematic review of the effect of home modification and fall prevention programs on falls and the performance of community-dwelling older adults. *Am J Occup Ther*. 2012 May-Jun;66(3):284-91.

Submitted 2-18-2022

VI.4 Evidence Based Medicine and Pregnancy (Janae McGuirk, GSM4)

In the United States alone there were approximately 3.6 million births in 2020 with a general fertility rate of 55.8 births per 1,000 women age 15 to 44¹. Most people know someone who has gone through pregnancy and childbirth and a majority of women will have experienced this process more than once. Women will frequently receive advice or recommendations from friends and family members that is either outdated, has not actually been clinically evaluated or can sometimes be dangerous. Additionally, if studies have been done related to pregnancy or childbirth, the actual recommendations can be overcautious, and may not give the individual the option of choosing what is right for them².

Although pregnancy is very common and affects a large percentage of the population; both directly and indirectly, there are sometimes challenges to finding evidence based medical recommendations for pregnant women. In addition to the challenge of finding proper sources, sometimes debunking incorrect information and delivering evidence-based recommendations can cause tension in the patient-doctor relationship or lead to misunderstanding, especially if not done in a considerate manner. The goal of this paper is to present some of the challenges of finding evidence-based resources for pregnant women and some possible solutions as well as how to present this information to the patient.

One of the greatest challenges in applying evidence-based research to pregnant populations is that pregnant women tend to be excluded from the vast majority of pharmacological, therapeutic, or preventive trials. The most compelling reason for this exclusion is due to fear of harm to the fetus, but also threat of legal liability, the complicated physiology of pregnant women, pregnant women being classified as a “vulnerable” population leading to the need for special protections in research and the vague wording of IRB regulations that tend to be interpreted conservatively for pregnant subjects. This exclusion can be detrimental to pregnant women who are not protected from being afflicted by different diseases or conditions that may benefit from medication. Additionally, approximately 64% of pregnant women are prescribed one or more medications during their pregnancy in the absence of adequate trials run on pregnant populations to justify the medication use³. Very few drugs are actually approved for use during pregnancy.

Other barriers to evidence based maternity care are lack of robust maternity performance measures and minimal commitment from primary stakeholders, perverse incentives of payment systems, loss of core childbearing knowledge and skills among health professionals and the most recent barrier of the prevalence of inaccurate information disseminated on social media and popular platforms. Although robust changes cannot be made overnight, some options for overcoming these barriers are: to increase awareness about deficits in the maternity care system and about evidence-based maternity care by educating and advising stakeholders; support research to further evidence-based maternity

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

care; reform the current reimbursement system to promote evidence-based maternity care; require performance measurement, reporting and improvement⁴.

The goal of evidence based maternity care is to use the best available research on the safety and effectiveness of specific practices to help guide maternity care decision-making and facilitate the best possible outcomes for mothers and newborns. Informed decision-making needs to be respectful of the values and circumstances of each individual woman, but also keep safety and effectiveness at the forefront. There are many practices that have become commonplace for pregnancy, labor and delivery, but lack sufficient data to support the intervention. An example of this is cesarean delivery. This procedure has become the most common surgery performed in the United States with over 1 million women delivering via C-section every year⁵. However, the rate of cesarean deliveries varies widely by state and even by different cities and towns within a state⁶. This is one of many interventions that although beneficial for the right populations, have become subject to the preferences and comfort level of different physicians or health systems instead of following strict evidence-based guidelines.

Ultimately, most pregnant women want to take responsibility for their own health and make choices base on informed advice. It is the job of the physician to provide accurate information and to help walk the patient through the validity of that information as well as the possible harms and benefits so each woman has the opportunity to individualize her care. Women will frequently encounter advice that lacks reason, evidence, or sufficient detail, but when the information is cited, and is evidence based, it carries its own validity and can help ease the stress of sorting through conflicting recommendations. When presenting evidence-based information to women about pregnancy, labor, and delivery, it is best to let them know the reasons behind recommendations and practice shared deision-making. This allows the patient (and physician) to honor their own preferences considering one's unique background, wishes, and goals for pregnancy and their child.

An example of the importance of high quality evidence based research is the PREMEVA study by Subtil, D. *et al*. They asked "Does treatment of bacterial vaginosis with clindamycin in pregnant women >18 y/o decrease late miscarriage or spontaneous very preterm birth?", in part because prior studies resulted in a lack of consensus about this approach. A high level of evidence is attainable when the principles of EBM are addressed, such as randomization, adequate calculated sample size, good follow-up, adequate duration, meaningful outcomes, demonstrated adherence, data and safety monitoring, and absence of conflicts, to name a few. This study screened 84,530 pregnant women before 14 weeks' gestation and allocated 2869 women with bacterial vaginosis to receive clindamycin or placebo. The authors showed no evidence of a reduction in risk of late miscarriage or spontaneous very preterm delivery after treatment with clindamycin. The results showed a relative risk increase of 10.1% (95% CI of 132% to -48%, NS) with $p = 0.82$, and number needed to treat of 951 (95% confidence interval of -118 to 116, NS)⁷. The authors concluded that there was little evidence that screening and

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

treating all pregnant women with bacterial vaginosis will prevent preterm delivery and its consequences when treatment begins before 20 weeks' gestation.

A high standard of inquiry and rigorous research methodology is needed in OB-GYN as in all disciplines. Clinical researchers and patients should expect no less. Understanding and practicing EBM is clearly a required skillset for all of us to embrace.

Citations

1. Hamilton, B., Martin, J., & Osterman, M. (2021). Births: Provisional data for 2020. *National Center for Health Statistics*. <https://doi.org/10.15620/cdc:104993>
2. Oster, E. (2021). *Expecting better: Why the conventional pregnancy wisdom is wrong--and what you really need to know*. Penguin Books.
3. Blehar, M. C., Spong, C., Grady, C., Goldkind, S. F., Sahin, L., & Clayton, J. A. (2013). Enrolling pregnant women: Issues in clinical research. *Women's Health Issues, 23*(1). <https://doi.org/10.1016/j.whi.2012.10.003>
4. Sakala, C., & Corry, M. P. (2008). *Evidence-based Maternity Care: What it is and what it can achieve*. Milbank Memorial Fund.
5. Sung S, Mahdy H. Cesarean Section. [Updated 2021 Dec 12]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK546707/>
6. CDC – National Center for Health Statistics – Homepage. https://www.cdc.gov/nchs/pressroom/sosmap/cesarean_births/cesareans.htm. April 15, 2022.
7. Subtil D, Brabant G, Tilloy E, Devos P, Canis F, Fruchart A, Bissinger MC, Dugimont JC, Nolf C, Hacot C, Gautier S, Chantrel J, Jousse M, Desseauve D, Plennevaux JL, Delaeter C, Deghilage S, Personne A, Joyez E, Guinard E, Kipnis E, Faure K, Grandbastien B, Ancel PY, Goffinet F, Dessein R. Early clindamycin for bacterial vaginosis in pregnancy (PREMEVA): a multicentre, double-blind, randomised controlled trial. *Lancet*. 2018 Nov 17;392(10160):2171-2179. doi: 10.1016/S0140-6736(18)31617-9. Epub 2018 Oct 12. PMID: 30322724.

Submitted 4-16-2022

VI.5 My experience with EBM and SGLT2 Inhibitors (Thomas Palladino, GSM 4)

Introduction

My experience with EBM and sodium-glucose cotransporter 2 inhibitors (SGLT2i) began during my third-year IM clerkship, when discussing treatment for patients with acute decompensated heart

failure. While most of the patients that I saw presented with heart failure with reduced ejection fraction (HFrEF; EF \leq 40%), there was significant discussion around a promising new option for patients with heart failure with preserved ejection fraction (HFpEF; EF >40%). Empagliflozin had recently been demonstrated to provide clinical benefit in patients with HFpEF, a notable observation for a drug initially developed to improve glycemic control in Type 2 diabetes through glucosuria. Later, I began my EBM elective interested in better understanding the current data available to support SGLT2i use in patients with heart failure. It was at this time that my first, general question arose:

How do SGLT2i, which act predominantly in the proximal tubule, exert cardiovascular benefits?

Before diving into trials, I explored some of the basic science around this topic. I came across several proposed cardioprotective mechanisms for SGLT2i.

SGLT2i appear to optimize left ventricular loading conditions. These agents block glucose and sodium reabsorption in the proximal tubule leading to osmotic diuresis and natriuresis, thereby reducing left ventricular preload. These agents reduce afterload via improved endothelial function, reduced aortic stiffness, and potentially V-gated K channel and protein kinase G-mediated vasodilation. There may be some improvements in cardiac metabolism and mechanical efficiency, potentially through myocardial utilization of ketones, production of which are known to increase due to SGLT2i. Further hypotheses for mechanisms underlying cardiovascular benefits from SGLT2i include reduction of cardiac fibrosis, improvement in the balance between pro-and anti-inflammatory signaling, and reduction in myocardial cytoplasmic sodium and calcium levels that have been implicated in experimental models of HF, as well as through an increase in sarcoplasmic calcium levels which would improve contractility. Further elucidation of these exact mechanisms will better delineate the role for SGLT2i in the heart failure population and its many subgroups. ¹

With this background in hand, I felt ready to begin reviewing trials. The EBM Database provided a useful starting point to gain an overview of landmark trials demonstrating cardiovascular benefits from SGLT2i. An EBM Database search for “SGLT” directed me to 6 entries. A search for “flozin” revealed two additional study summaries, and some more targeted searches directed me towards specific trials that did not have one of these phrases in their title. I landed on a fundamental question:

How did antihyperglycemic SGLT2i become such a robust area of interest in cardiovascular medicine?

The EMPA-REG Outcome study from 2015 seemed like a reasonable place to start. This trial sought to assess cardiovascular benefits from SGLT2i, given their known glucose-lowering effects and postulated beneficial effects on vascular health among other cardiovascular mechanisms. The study included patients with Type 2 Diabetes and established cardiovascular disease, and randomized patients to either 10mg empagliflozin, 25mg empagliflozin, or placebo daily. Data from both empagliflozin groups was pooled for analysis and showed significant risk reduction for the primary composite end point of

CV death, nonfatal MI, or nonfatal stroke (RRR 13%; 95% CI [0.7% to 24.6%]; NNT 61; 95% CI [1267 to 31]), as well as for the secondary outcome which was a composite of the primary outcome components along with hospitalization for unstable angina (RRR 11%; 95% CI [1.4% to 20.9%]; NNT 66; 95% CI [534 to 31]). When assessed individually, the following were significantly reduced in the treatment group: death from any cause (RRR 31%; 95% CI [17.5% to 42.2%]; NNT 38; 95% CI [76 to 25]), CV death (RRR 37.5%; 95% CI [22.2% to 49.8%]; NNT 45; 95% CI [87 to 30]), and hospitalization for heart failure (RRR 34%; 95% CI [14.2% to 49.2%]; NNT 72; 95% CI [201 to 42]). Rates of fatal or nonfatal MI and fatal or nonfatal stroke were not significantly different between groups. Notably, empagliflozin was associated with risk of genital infection (6.4% of empagliflozin patients vs. 1.8% of placebo patients; $p < 0.01$). This trial's patients had high cardiovascular risk at baseline, and the study demonstrated cardiovascular benefit from SGLT2i, potentially through mechanisms other than lowering glucose alone.²

The authors did note potential concern about renal safety, so I next focused on results from a new trial recommended to me through the ACCESS platform, to which I had subscribed to receive updates about pre-appraised evidence in my areas of interest.

How do SGLT2i impact those with CKD?

The EMPA-Kidney study expanded upon previous studies of SGLT2i to demonstrate significant risk reduction for the composite outcome of progression of kidney disease or death from cardiovascular causes in patients with varied baseline GFR, degree of albuminuria, and CKD etiology, regardless of diabetes status (RRR 23%; 95% CI [13% to 31%]; NNT 26; 95% CI [48 to 18]). This was driven by progression of CKD, as when assessed individually, this component outcome was significantly reduced in the treatment group (RRR 24%; 95% CI [13.8% to 32.7%]; NNT 28; 95% CI [50 to 19]) but CV death was not (RRR 14%; 95% CI [-20.7% to 39.4%]). There were no significant differences in safety outcomes between treatment groups including serious UTIs, genital infections, hyperkalemia, AKI, or dehydration.³

The CREDENCE⁴ trial showed SGLT2-mediated benefit on the composite primary outcome of kidney disease progression and cardiovascular death with canagliflozin (RRR 28%; 95% CI [16.1% to 38.3%]; NNT 23; 95% CI [43 to 16]). Similarly, DAPA-CKD⁵ demonstrated benefit from dapagliflozin in reducing the risk of the composite outcome of decline in eGFR $\geq 50\%$ and ESRD (RRR 42%; 95% CI [28.8% to 52.1%]; NNT 21; 95% CI [33 to 16]) as well as both individual outcomes. However, these trials included a predominance of patients with CKD secondary to diabetes with increased albuminuria. The majority of CKD patients overall do not have diabetes and have lower levels of albuminuria, so it is notable that EMPA-Kidney established benefit in patients more representative of the overall CKD population.

Now that I had this general background on SGLT2i and cardiovascular benefits as well as safety and efficacy in CKD, I circled back to my original question:

Benefit from pharmacologic therapy has been elusive in HFpEF. What is the evidence for cardiovascular benefit from SGLT2i in patients with HFpEF?

The EBM Database had an entry that examined this exact question. The EMPEROR-Preserved Trial studied daily empagliflozin in patients with NYHA Class II-IV HFpEF and showed a significant reduction in the composite outcome of CV death and HF hospitalizations in the treatment group (RRR 19%; 95% CI [9.1% to 28.4%]; NNT 30; 95% CI [19 to 68]). When the components of the composite were assessed individually, hospitalizations were significantly lower in the treatment arm (RRR 27%; 95% CI [15.1% to 37.4%]; NNT 31; 95% CI [21 to 60]), but CV deaths were not (RRR 9%; 95% CI [-6.0% to 25.3%]). The effect of empagliflozin was consistent regardless of diabetes status, and rates of serious adverse events were similar between groups.⁶

This study followed several trials examining SGLT2i in patients with HFrEF, so I turned my attention towards these next:

What is the evidence for cardiovascular benefit from SGLT2i in patients with HFrEF?

The EMPEROR-Reduced⁷ and DAPA-HF⁸ trials addressed this question, and summaries were available in the EBM Database. EMPEROR-Reduced studied daily empagliflozin in patients with NYHA Class II-IV HFrEF and showed a significant reduction in the composite outcome of CV death and HF hospitalizations in the treatment group (RRR 22%; 95% CI [11.6% to 30.7%]; NNT 19; 95% CI [37 to 12]). Like in EMPEROR-Preserved, this was driven by a reduction in HF hospitalizations (RRR 28%; 95% CI [16% to 38%]; NNT 20; 95% CI [36 to 13]), as when assessed alone, there was no significant difference between groups in terms of CV death (RRR 7.2%; 95% CI [-12 to 23]). Again, there was significant risk reduction in the composite primary outcome regardless of diabetes status. Uncomplicated genital tract infections were more frequent in the empagliflozin group (1.7%) than the placebo group (0.6%). DAPA-HF studied daily dapagliflozin in patients with NYHA II-IV HFrEF and showed a significant reduction in the composite outcome of CV death, HF hospitalizations, or HF urgent visits, regardless of diabetes status (RRR 23%; 95% CI [13.4% to 31.9%]; NNT 20; 95% CI [37 to 14]). This was driven by significant risk reduction for all three of these outcome components: CV death (RRR 17%; 95% CI [1.9% to 29.7%]; NNT 51; 95% CI [501 to 27]), HF hospitalizations (RRR 27%; 95% CI [14.9% to 38.1%]; NNT 27; 95% CI [54 to 18]), and HF urgent visits (RRR 57%; 95% CI [8.9% to 79%]; NNT 182; 95% CI [1416 to 94]). There was no difference in adverse events between groups, including volume depletion, renal events, hypoglycemia, Fournier's gangrene, amputation, fracture, and ketoacidosis.

These three trials showed similar results for patients with stable, chronic HFpEF and HFrEF, regardless of diabetes status. I then focused on a different population:

What is the evidence for cardiovascular benefit from SGLT2i in patients with acute HF?

I found an entry on the SOLOIST-WHF Trial in the EBM Database, which showed sotagliflozin's significant risk reduction in the composite outcome of CV deaths, HF hospitalizations, and HF urgent

visits for diabetic patients during or shortly after an episode of acute decompensated heart failure, regardless of EF subgroup (RRR 30.3%; 95% CI [21.6% to 38.1%]; NNT 6; 95% CI [8 to 4]). HF hospitalizations and HF urgent visits (RRR 34%; 95% CI [24% to 42.8%]; NNT 6; 95% CI [9 to 5]) but not CV deaths (RRR 11.2%; 95% CI [-27.2% to 38%]) were significantly reduced in the treatment group when assessed individually. Rates of diarrhea, genital mycotic infections, urinary tract infections, and volume depletion were higher in the sotagliflozin group, but not significantly different from those in the placebo group.⁹

I then found the EMPULSE trial via PubMed, which showed significant benefit from empagliflozin via a stratified win ratio with a hierarchical composite of death, number of total HF events, time to first HF event, and change in Kansas City Cardiomyopathy Questionnaire-Total Symptom Score (KCCQ-TSS) from baseline to 90 days (Win Ratio 1.36; 95% CI [1.09 to 1.68] p=0.0054). The population studied was patients with stable, acute de novo or decompensated HF regardless of EF or diabetes status. The population differed from SOLOIST's in that it included patients with acute de novo HF. This study's demonstrated benefit of adding SGLT2i to traditional first-prescribed guideline-directed medical therapy agents is notable, as seen in the de novo group who would not have had prior heart failure treatments (ACE-inhibitors/ angiotensin receptor blockers/angiotensin receptor-neprilysin inhibitors, mineralocorticoid receptor antagonists, beta blockers). EMPULSE showed no differences between treatment groups in terms of worsening renal function, volume depletion, or ketoacidosis, but an increase in Hgb and Hct that likely reflected the diuretic response to empagliflozin.¹⁰

Taken together, the results from these trials support expanding the role for SGLT2i to the acute and post-acute windows. SGLT2i in these settings had previously not been studied, except in the EMPA-RESPONSE-AHF trial, which was a pilot study that suggested clinical benefit from empagliflozin in a small cohort of patients (n=80) with acute heart failure.¹¹ Acute HF involves dynamic fluid, electrolyte, and hemodynamic changes, and the first several months after discharge are also a vulnerable time. These trials demonstrated that empagliflozin is safe and effective in these periods.

Now that I had explored HFpEF, HFrEF, and acute HF, I wondered about those patients who fell somewhere in between these groups.

What is the evidence for cardiovascular benefit from SGLT2i in patients who previously had HFrEF, now with recovered EF >40%?

I found the recently published DELIVER trial via PubMed, which showed benefit from dapagliflozin in reducing risk of the composite outcome of worsening HF or CV death in patients with and without Type 2 diabetes with NYHA class II-IV heart failure with EF >40%, as well as those previously with EF <40% that had recovered to ≥40% by enrollment. The primary outcome benefits were significant across EF subgroups (RRR 16%; 95% CI [6.6% to 24.5%]; NNT 32; 95% CI [82 to 20]). Additionally, the dapagliflozin group had a significantly higher change from baseline in KCCQ scores, reflecting reduction in symptom burden (Win Ratio, 1.11; 95% CI [1.03 to 1.21] p=0.009). Rates of adverse events were

similar between groups. This trial included patients who were hospitalized or recently hospitalized, as well as those with recovered EF, which was a broader group than previously studied. Patients with recovered EF are often excluded from clinical trials but represent a growing population due to greater success in treating heart failure with reduced EF, so it is important that dapagliflozin appears to benefit this group.¹²

At this point, I wanted to revisit an interesting point that piqued my interest while reading the Database entry on sotagliflozin, the only agent discussed in this chapter that provides mixed SGLT1/2 blockade.

What is the effect of combined SGLT1/2 inhibition from sotagliflozin?

The SCORED Trial¹² demonstrated that sotagliflozin reduced the composite risk of cardiovascular death, HF hospitalizations, and HF urgent visits in patients with Type 2 diabetes, CKD regardless of albuminuria, and additional CV risk factors across subgroups (RRR 24%; 95% CI [14.6% to 33.3%]; NNT 41; 95% CI [73 to 28]), but was notably associated with numerous adverse events. Diarrhea (RRI 42%; 95% CI [24% to 63%]; NNH 40; 95% CI [29 to 66]), genital mycotic infections (RRI 178%; 95% CI [289% to 978%]; NNH 67; 95% CI [50 to 96]), volume depletion (RRI 30.4%; 95% CI [9.5% to 55.2%]; NNH 82; 95% CI [49 to 236]), diabetic ketoacidosis (RRI 114%; 95% CI [13.7% to 303%]; NNH 331; 95% CI [177 to 1804]), and hypotension (RRI 37.9%; 95% CI [6.9% to 77.8%]; NNH 140; 95% CI [78 to 664]) were significantly associated with sotagliflozin. In addition to the proximal tubule, SGLT1 is expressed in the small intestinal brush border, and inhibition of the transporter lowers postprandial glycemia via slowing intestinal glucose absorption; the authors suggest that the increased rates of diarrhea in the treatment arm may be due to SGLT1's GI expression. When studied in Type 1 diabetes patients, sotagliflozin significantly lowered blood glucose and promoted weight loss without significantly increasing hypoglycemic events, but the rates of ketoacidosis were 5 times as high as seen in the placebo group.¹⁴ It is unclear whether SGLT1 inhibition provides additional cardioprotective benefit and if a plausible mechanism exists, but this may be worth further investigation.

While the composite primary end point showed significant benefit from sotagliflozin, this appears to be largely driven by HF hospitalizations and HF urgent visits, as there was no significant risk reduction in CV deaths.

The evidence for cardiovascular benefit from SGLT2i is robust, but let's revisit potential harms.

While some of the trials reviewed showed no difference in rates of adverse events between groups, SCORED made it clear that these agents do not come without risks. Rates of diarrhea, genital mycotic infections, volume depletion, diabetic ketoacidosis, and hypotension were significantly higher in the sotagliflozin group. As above, 95% CIs for the RRI were variable and sometimes quite large, and NNH ranged from 40-331 with variable CIs. Diarrhea may be attributable to SGLT1 blockade and the resulting osmotic load from increased intestinal glucose. Empagliflozin was associated with genital

infections in EMPA-REG Outcome and EMPEROR-Reduced, likely owing to the favorable microbial environment created by glucose-rich urine.

Both hyperglycemic and euglycemic ketoacidosis are uncommon but previously documented adverse effects of SGLT2i. Although event numbers were too small to calculate hazard ratios in EMPA-Kidney, there were 6 episodes of ketoacidosis in the empagliflozin group (one in a patient without diabetes) and one in the placebo group. Euglycemic ketoacidosis is a unique entity that warrants special attention. The mechanism has been postulated to involve the significant SGLT2i-mediated increase in renal glucose clearance outpacing endogenous glucose production, resulting in elevated catecholamine and corticosterone production that then drives lipolysis. As euglycemic ketoacidosis may not often be encountered, clinicians should be aware of this risk as SGLT2i use expands.^{15, 16}

Now, how to put it all together?

Overall, SGLT2i are a promising class that have demonstrated clinical benefit in numerous populations, including patients with: chronic HFrEF, chronic HFpEF, acute decompensated HF, acute de novo HF, CKD regardless of albuminuria or eGFR, as well as those with and without diabetes in the above categories. The significant results, in general, showed RRR between 20s%-40s% with reasonably tight CIs. Significant primary end points were largely driven by reductions in HF hospitalizations and/or urgent visits but not CV deaths – EMPA-REG and DAPA-HF are two exceptions. The fortuitous discovery of cardiovascular benefits observed with these agents originally developed to treat Type 2 diabetes is a welcome development in the realm of HF therapy, but further investigation is warranted. A better understanding of the basic mechanisms underlying these clinical benefits as well as longer follow-up with adequately powered subgroup analyses will be important to better define the role for SGLT2i in the broad spectrum of HF patients, and to determine if there is a true mortality benefit for some. Most studies had a limited number of patients on angiotensin receptor-neprilysin inhibitors (ARNIs), which is a relatively new class of medications. These results will be worth revisiting as use of ARNIs expands, which may affect the relative benefit from SGLT2i. Importantly, the adverse event profile is not insignificant. As above, longer follow-up with adequate sample size will be important in understanding the true scope and severity of these events, and how to best balance their risk with potential cardiovascular benefit.

I chose to focus my EBM searches on trials that included important clinical endpoints such as death or hospitalizations, as opposed to relying solely on surrogate endpoints like echocardiographic or laboratory data, or quality of life questionnaires. So naturally, this chapter is not a totally exhaustive review of the evidence around SGLT2i and cardiovascular outcomes. The figure below includes a useful summary of the body of work in the HF space, including ongoing trials as well.

How has this experience impacted me?

I had the opportunity to start with a question that arose in the clinical setting and explore the breadth of available evidence to answer that question. I became better versed in efficiently identifying and

Evidence Based Medicine Study Guide
 EBM Elective
 Department of Medicine

reviewing clinical trials of interest, understanding and interpreting their outcome measures, as well as critiquing aspects of study design and methodology. I allowed myself to meander from question to question as they arose, while maintaining a systematic approach to evaluating trials and reviews. I plan to adopt a version of this process as I continue training in internal medicine, where clinical questions around promising new therapies will no doubt arise. Additionally, I have gained a useful fund of knowledge on the potential benefits and drawbacks of SGLT2i that will certainly be useful as they become more widely prescribed.

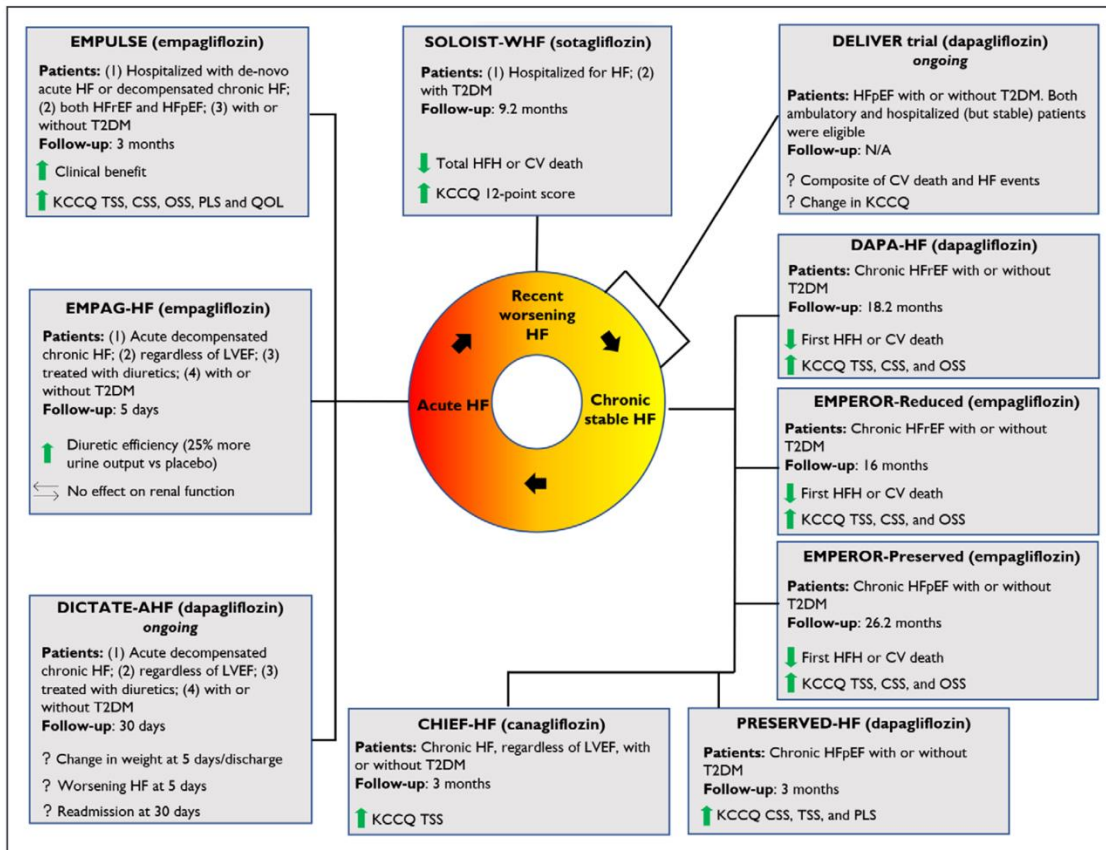


Figure 1. An overview of completed and ongoing studies of SGLT2i in the full spectrum of HF populations¹⁷

References

1. Verma, S., & McMurray, J. J. (2018). SGLT2 inhibitors and mechanisms of cardiovascular benefit: a state-of-the-art review. *Diabetologia*, 61(10), 2108-2117

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

2. Zinman, B., Wanner, C., Lachin, J. M., Fitchett, D., Bluhmki, E., Hantel, S., ... & Inzucchi, S. E. (2015). Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. *New England Journal of Medicine*, 373(22), 2117-2128.
3. EMPA-KIDNEY Collaborative Group. (2022). Empagliflozin in Patients with Chronic Kidney Disease. *New England Journal of Medicine*.
4. Perkovic, V., Jardine, M. J., Neal, B., Bompoint, S., Heerspink, H. J., Charytan, D. M., ... & Mahaffey, K. W. (2019). Canagliflozin and renal outcomes in type 2 diabetes and nephropathy. *New England Journal of Medicine*, 380(24), 2295-2306.
5. Heerspink, H. J., Stefánsson, B. V., Correa-Rotter, R., Chertow, G. M., Greene, T., Hou, F. F., ... & Wheeler, D. C. (2020). Dapagliflozin in patients with chronic kidney disease. *New England Journal of Medicine*, 383(15), 1436-1446.
6. Anker, S. D., Butler, J., Filippatos, G., Ferreira, J. P., Bocchi, E., Böhm, M., ... & Packer, M. (2021). Empagliflozin in heart failure with a preserved ejection fraction. *New England Journal of Medicine*, 385(16), 1451-1461.
7. Packer, M., Anker, S. D., Butler, J., Filippatos, G., Pocock, S. J., Carson, P., ... & Zannad, F. (2020). Cardiovascular and renal outcomes with empagliflozin in heart failure. *New England Journal of Medicine*, 383(15), 1413-1424.
8. McMurray, J. J., Solomon, S. D., Inzucchi, S. E., Køber, L., Kosiborod, M. N., Martinez, F. A., ... & Langkilde, A. M. (2019). Dapagliflozin in patients with heart failure and reduced ejection fraction. *New England Journal of Medicine*, 381(21), 1995-2008.
9. Bhatt, D. L., Szarek, M., Steg, P. G., Cannon, C. P., Leiter, L. A., McGuire, D. K., ... & Pitt, B. (2021). Sotagliflozin in patients with diabetes and recent worsening heart failure. *New England Journal of Medicine*, 384(2), 117-128.
10. Voors, A. A., Angermann, C. E., Teerlink, J. R., Collins, S. P., Kosiborod, M., Biegus, J., ... & Ponikowski, P. (2022). The SGLT2 inhibitor empagliflozin in patients hospitalized for acute heart failure: a multinational randomized trial. *Nature medicine*, 28(3), 568-574.
11. Damman, K., Beusekamp, J. C., Boersma, E. M., Swart, H. P., Smilde, T. D., Elvan, A., ... & Voors, A. A. (2020). Randomized, double-blind, placebo-controlled, multicentre pilot study on the effects of empagliflozin on clinical outcomes in patients with acute decompensated heart failure (EMPA-RESPONSE-AHF). *European journal of heart failure*, 22(4), 713-722.
12. Solomon, S. D., McMurray, J. J., Claggett, B., de Boer, R. A., DeMets, D., Hernandez, A. F., ... & Langkilde, A. M. (2022). Dapagliflozin in heart failure with mildly reduced or preserved ejection fraction. *New England Journal of Medicine*, 387(12), 1089-1098.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

13. Bhatt, D. L., Szarek, M., Pitt, B., Cannon, C. P., Leiter, L. A., McGuire, D. K., ... & Steg, P. G. (2021). Sotagliflozin in patients with diabetes and chronic kidney disease. *New England Journal of Medicine*, 384(2), 129-139.
14. Rendell, M. S. (2018). Sotagliflozin: a combined SGLT1/SGLT2 inhibitor to treat diabetes. *Expert Review of Endocrinology & Metabolism*, 13(6), 333-339.

VI.4 Evidence Based Medicine and Substance Use Disorders (Nikki Ratnapala, GMS4)

Introduction

People with substance use disorders (SUD) are diverse, with wide variations in terms of substances used, comorbidities, and psychosocial complexities. This varied population of people pose unique challenges when studying SUD, which are imperative to consider when designing studies. Through the use of high quality studies, therapies have been shown to address specific types of SUD, including alcohol, opiates, cocaine, and tobacco. Treatments have also been developed to address not only the substance use, but also the array of issues that commonly coincide with SUD, including fractured relationships, legal issues, employment, and co-occurring medical and psychiatric disorders. In this chapter I aim to highlight some challenges to studying this population and a small portion of existing data published on those with SUD.

Challenges in Studying Substance Use Disorders

Throughout the world, SUD continues to pose a significant risk to public health. It is imperative that researchers direct attention to ways in which the medical community can help reduce the pain and suffering caused by this group of disorders. However, the inclusion of those with SUD, often considered a vulnerable population, as participants in research studies presents several increased risks and unique challenges that need to be addressed in study protocols.

Risks to People with Substance Use Disorders Participating in Research

Major ethical challenges exist for substance use research, with many of these challenges unresolved. Issues exist in many areas, including ability to provide consent, confidentiality concerns, legal considerations, and researcher understanding of the political, social, and economic setting in which they work¹. On the issue of consent, informed consent requires the subject to both comprehend the risks and benefits and willingly volunteer. As SUD research oftentimes includes those who are intoxicated or undergoing detox, there is concern that this population of subjects are unable to fully give consent. One interesting way researchers addressed this issue was in a study investigating efficacy of intranasal vs. intramuscular naloxone during an overdose, which consented subjects at a safe injection site before subjects injected². This mitigated some of the concerns with ability to provide consent when subjects would be unable to do so. Regarding confidentiality, the US has protections for those who partake in research, a Certificate of Confidentiality³, but it is not the standard worldwide. Of note, unlike physician-patient and attorney-client relationships, the researcher-participant relationship is not privileged and is not given the same protections. As many types of substance use are illegal, and are associated with illegal activities (such as driving while intoxicated, or selling illegal substances), it is imperative to consider participants' current legal issues when conducting research.

Substance Use Disorder Population Challenges

Longitudinal studies are essential to measuring outcomes. Barriers to obtaining longitudinal data on those with SUD include unstable housing and income, difficulty obtaining up to date contact information, transportation difficulties, and a lack of understanding of the communities in which

subjects are a part. It is important to consider the increased likelihood that populations with SUD have an increased unstable lifestyle overall compared to the general population, sometimes making it difficult for participants to track and make it to appointments and thus partake in longitudinal studies.

Due to these issues, retention rates in studies which include those with SUD can be expected to be low, and thus affect study findings. One recent meta-analysis of 151 studies on SUD found an average dropout rate of 30%⁴. This number is in line with the dropout rates which I personally observed when reviewing SUD research. However, research is ongoing on ways to increase follow-up rates within SUD populations^{5,6}. There are also published resources which aim to increase follow-up rates within this population⁷. Notable themes from this ongoing research and resources include the importance of rapport building; obtaining contacts from the participant such as family and friends who have a stable lifestyle, and thus would be more likely to get in contact with the participant; use of text message reminders; graduated incentives; and using trackers. One such tracker was used in a study I reviewed for the EBM database which used the timeline follow-back method to report alcohol consumption⁸, a validated calendar-based method of self-reported use of substances⁹.

In order to successfully study SUD, it is imperative for researchers to address these unique issues. This will involve continuing to study ways in which participant retention is improved, as well as ensuring potential subjects of an SUD study understand what participation looks like in the short and long term. It has been emphasized that researchers must plan for ways to address common concerns within the SUD population at the time of enrollment and throughout study periods (eg, providing transportation reimbursement and including study information on social media or an easily accessible website so participants can easily get in touch with researchers)¹⁰.

Study Designs: Randomized Control Trials vs. Others

Randomized controlled trials (RCTs) are crucial to test and develop new therapies for SUD. However, as discussed above, there remain social, ethical, and logistical challenges to conducting studies, especially longitudinal, among those with SUD. As the medical community has shed light on importance of growing SUD research¹¹, it is of utmost importance to find ways to successfully complete RCTs within SUD populations. One such way is to complete different types of studies, such as qualitative, cohort, or systematic reviews, and use findings to inform future RCTs testing therapies for SUDs.

For instance, researchers could help improve the design and implementation of future RCTs, as those who work directly with participants with SUD often learn of unique features that can help in study optimization. Qualitative interviews in particular allow for feedback from participants that may highlight ways to improve recruitment, protocol adherence and retention. Cohort studies are particularly helpful when studying SUD. In fact, I was tempted to review a large cohort study on methadone vs. buprenorphine in pregnant persons with opioid use disorder¹², and even talked a reluctant Dr. Ross into adding it to the EBM database. Before I reveal why I decided not to, I want to tell you why it remains an important publication.

It is difficult to study SUD populations. It is even more difficult to study outcomes in those who have SUD and are pregnant. This large cohort study was able to show an association of lower risk of adverse

neonatal outcomes in those who took buprenorphine versus methadone. They had impressive numbers (n= 2,548,372), were able to perform an analysis with data that already existed, and thus was far less costly than a RCT, and the results of the study will likely help guide future clinical management. However, cohort studies do not prove a causal relationship- they are a second-best option to RCTs. This is the reason I decided to instead add an RCT comparing methadone vs. buprenorphine in pregnant persons with opioid use disorder. That trial of 131 neonates revealed mothers in the buprenorphine-treated group required significantly less morphine, (mean dose 1.1 mg vs 10.4 mg), had shorter hospital stays (10 days vs 17.5 days), and had a significantly shorter duration of treatment for neonatal abstinence syndrome (4.1 days vs 9.9 days), all highly statistically significant at $p < 0.009$ ¹³. Even though the number of subjects pales in comparison, this study has stronger data due to the fact it is a RCT. Bottom line: RCTs remain the gold standard to test and develop new therapies, and future research must use novel means to make successful RCTs within SUD populations possible. In the following section, I review some of results of literature searches revealed concerning treatment for SUD. It is by no means a comprehensive review of evidence-based data on those with SUD, but rather a highlight of evidence I found interesting and/or added to the EBM database, concerning pharmacotherapy and psychosocial and behavioral interventions.

Pharmacotherapy

There exists a body of evidence-based pharmacotherapies for SUD, including for alcohol use disorder, tobacco use disorder, and opioid use disorder.

- a. Naltrexone is now the most commonly used medication to treat alcohol use disorder, and was first shown to be effective within this population over 30 years ago¹⁴. While not as commonly used as naltrexone, gabapentin has been shown to be effective in treating alcohol use disorder. In one trial, authors found a larger number of gabapentin-treated individuals had no heavy drinking days (27% versus 9%), a difference of 18.6% (95% CI, 3.1-34.1; $P = .02$; NNT 5.4), and more total abstinence compared with placebo (18% versus 4%), a difference of 13.8% (95% CI, 1.0-26.7; $P = .04$; NNT, 6.2). This suggests that gabapentin may be efficacious in preventing relapse to heavy drinking and in promoting abstinence in patients with a history of alcohol withdrawal symptoms¹⁵.

Another very interesting and exciting pharmacotherapy in the treatment of alcohol use disorder is the use of hallucinogens. One study investigated if psilocybin administered in combination with psychotherapy decreased the percentage of heavy drinking days versus placebo plus psychotherapy. In this trial, authors found the percentage of heavy drinking days during the 32 weeks was 9.7% for the psilocybin group and 23.6% for the placebo (diphenhydramine) group, a mean difference of 13.9%; (95% CI, 3.0-24.7; $P = .01$)⁸. While other RCTs have been published in the 1970s showing efficacy of LSD in treatment of alcohol use disorder,¹⁶ this was the first RCT of psilocybin for alcohol use disorder.

- b. Nicotine replacement therapy (patch plus gum or lozenge) is the first-line treatment of tobacco use. There is current data that show varenicline to be the most efficacious form of pharmacotherapy for tobacco use disorder¹⁷. Cytisine was thought to be a promising addition to pharmacotherapy for tobacco use disorder. In one trial, researchers found that at 1 month, continuous abstinence from smoking was higher in participants receiving cytisine versus nicotine-replacement therapy (40% vs. 31%), a difference of 9.3% (CI 4 to 14), a NNT 11 (CI 7 to 24), $P < 0.001$, with sustained findings at 6 months (22% vs. 15%, $P = 0.002$). However, self-reported adverse events over 6 months occurred more frequently in the cytisine group (31% vs. 20%, NNH 9 to 17)¹⁸. In addition, in a follow-up study, cytisine did not meet noninferiority versus varenicline, and is not currently used in the US for tobacco use disorder¹⁹.
- c. Methadone pharmacotherapy has been shown efficacy in treating opioid use disorder since the 1960s²⁰. It remains as one of the most commonly used medications to treat opioid use disorder, along with buprenorphine. Data have shown that buprenorphine alone and in combination with naloxone reduce the use and craving for opioids versus placebo. In that trial, the percent of urine samples negative for the combination was 17.8% compared to 5.8% for placebo, a number needed to treat (NNT) of 8 (CI 5 to 29), and of buprenorphine alone was 20.7% compared to 5.8% for placebo, a NNT of 6 (CI 4 to 15). Although the study was small, and stopped early due to demonstrated efficacy, it influenced the approval by the FDA of this now commonly prescribed therapy²¹. Buprenorphine, as opposed to methadone, has the benefit of having a longer half-life, with differing formulations including sublingual, subcutaneous, and via implant. In one study, among opioid-dependent adults maintaining clinical stability, subjects randomized to buprenorphine implants were found to have higher rates of no opioid use for at least 4 of 6 months studied compared to sublingual buprenorphine (96% implant versus 88% sublingual, CI 0.009 to ∞ , $P < .001$)²². Despite the fact that data from this study influenced the approval of the subdermal implant by the FDA in May 2016, it was discontinued due to multiple factors, including the delivery system, reimbursements, and inability to commercialize. Data has also shown that weekly and monthly subcutaneous buprenorphine are noninferior to sublingual buprenorphine within a population of those with opioid use disorder²³. In that trial, the percent of subjects without evidence of opioid use for the subcutaneous group was 17.4% versus 14.4% for the sublingual group, a RRI 21% (95% CI, 86.7 to -22.3; $P = \text{NS}$). Additionally, the percent of all urine samples negative for opioids was 35.1% for the subcutaneous group versus 28.4% for the sublingual group, a NNT of 15 (CI 11 to 21). The findings from this trial helped influence FDA approval of subcutaneous buprenorphine, and it is now commonly used as maintenance therapy for treatment of opioid use disorder.

Psychosocial/Behavioral Interventions

Many of the above referenced publications also included psychosocial and/or behavioral interventions, as holistic treatment for SUD often include such therapy. It has been shown that best practices in addiction treatment should include pharmacotherapy plus behavioral therapy, based on a metanalysis

of 30 RCTs that included +/- pharmacotherapy +/- behavioral therapy²⁴. Psychosocial interventions have also been shown as an effective adjunct to treating SUD. One such study found that providing personalized patient navigation that included proactive, personalized services including barrier resolution, motivational intervention, support and encouragement, advocacy with other providers, and linkage to resources for basic needs (such as food, housing, clothing, and transportation) was effective in reducing hospital readmissions and ED visits among previously hospitalized patients with comorbid substance use disorders²⁵. In that trial, inpatient admissions per 1000 person-days were 6.05 for the patient navigation group versus 8.13 for usual care group, a hazard ratio of 0.74 (CI 0.58 to 0.96; P= 0.020). ED visits per 1000 person-days were 17.66 for the patient navigation group versus 27.85 for usual care group, a hazard ratio of 0.66 (CI 0.49 to 0.89; P= 0.006). In a follow-up to this study, authors found that the patient navigation group generated \$17,780 per participant in cost savings, underscoring the importance of psychosocial interventions can have on the future of our health system²⁶.

Bottom line: How good are our current treatments? Does treatment really work?

The answer to this question is aggravatingly simple: it varies. Treatments for SUD are an array of different approaches to this diverse population. Response to treatment is not “yes” or “no”, but rather of gradations of improvement. Additionally, research settings are not perfectly controlled experiments, but complex and challenging in real lives. To be clear, the existing guidelines for treating SUD are effective for some, as the referenced data above has shown. However, this also brings up the complexity of how we measure “good”. Just because a trial shows efficacy or non-inferiority, does not necessarily mean the treatment under study is “good”, if we measure “good” as working for most. A cross-sectional study on a population of persons with SUD found that those who used treatment and/or recovery services were significantly associated with a fewer number of recovery attempts, as defined by self-reporting recovery from SUD²⁷. The study also found that the median number of relapses before recovery was two. This suggests that what treatments we have available work, but it takes more than one attempt to recover and to remain substance-free over time. While current research has led to improvements in treatments for SUD, there remains much to improve.

Personal Takeaway & Reflection

My interest in SUD intensified during my addiction psych rotation at the VA. It was here that I was not only inspired by the determination and perseverance of those recovering from SUD, but also came to appreciate how a holistic approach to treating people with SUD is required. In this vein, I became interested in the evidence the medical community has gathered to better serve this unique population of people. Throughout the EBM course, I was given space to move from a clinician team member role, to a role of critically appraising current evidence, to a role where I brainstormed more as an addiction researcher; a role I am interested in stepping into in the future. Particularly while writing this EBM chapter, I thought more on the gaps in knowledge we have on addiction and SUD, the ways such gaps may be filled, and how research into better ways to study this oftentimes challenging population is likely

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

required to promote and achieve healthier outcomes. What gives me particular hope is how the mindset around addiction is changing for the better- it is now more often treated as it should be, a disease, as opposed to a moral fracture. In fact, the US's general opinion of substances and how we can both use them and treat them has undergone large changes. A great example is an article in the New York Times published on the day I am writing this (1/3/23), which sheds light on use of psilocybin, which was legalized in Oregon for use, under trained supervision, for treatment of psychiatric disorders, including alcohol use disorder²⁸. This spotlight on finding new ways to treat SUD has me excited for the future of this important research.

References

1. Ryan JE, Smeltzer SC, Sharts-Hopko NC. Challenges to Studying Illicit Drug Users. *J Nurs Scholarsh*. 2019;51(4):480-488.
2. Dietze P, Jauncey M, Salmon A, et al. Effect of Intranasal vs Intramuscular Naloxone on Opioid Overdose: A Randomized Clinical Trial. *JAMA Netw Open*. 2019;2(11):e1914977.
3. Wolf LE, Beskow LM. New and Improved? 21(st) Century Cures Act Revisions to Certificates of Confidentiality. *Am J Law Med*. 2018;44(2-3):343-358.
4. Lappan SN, Brown AW, Hendricks PS. Dropout rates of in-person psychosocial substance use disorder treatments: a systematic review and meta-analysis. *Addiction*. 2020;115(2):201-217.
5. Bricca A, Swithenbank Z, Scott N, et al. Predictors of recruitment and retention in randomized controlled trials of behavioural smoking cessation interventions: a systematic review and meta-regression analysis. *Addiction*. 2022;117(2):299-311.
6. Klimas J, Hamilton MA, Gorfinkel L, Adam A, Cullen W, Wood E. Retention in opioid agonist treatment: a rapid review and meta-analysis comparing observational studies and randomized controlled trials. *Syst Rev*. 2021;10(1):216.
7. Scott CK. A replicable model for achieving over 90% follow-up rates in longitudinal studies of substance abusers. *Drug Alcohol Depend*. 2004;74(1):21-36.
8. Bogenschutz MP, Ross S, Bhatt S, et al. Percentage of Heavy Drinking Days Following Psilocybin-Assisted Psychotherapy vs Placebo in the Treatment of Adult Patients With Alcohol Use Disorder: A Randomized Clinical Trial. *JAMA Psychiatry*. 2022;79(10):953-962.
9. Hjorthoj CR, Hjorthoj AR, Nordentoft M. Validity of Timeline Follow-Back for self-reported use of cannabis and other illicit substances--systematic review and meta-analysis. *Addict Behav*. 2012;37(3):225-233.
10. Polcin DL, Mericle A, Callahan S, Harvey R, Jason LA. Challenges and Rewards of Conducting Research on Recovery Residences for Alcohol and Drug Disorders. *J Drug Issues*. 2016;46(1):51-63.
11. Lembke A, Humphreys K. The Opioid Epidemic as a Watershed Moment for Physician Training in Addiction Medicine. *Acad Psychiatry*. 2018;42(2):269-272.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

12. Suarez EA, Huybrechts KF, Straub L, et al. Buprenorphine versus Methadone for Opioid Use Disorder in Pregnancy. *N Engl J Med.* 2022;387(22):2033-2044.
13. Jones HE, Kaltenbach K, Heil SH, et al. Neonatal abstinence syndrome after methadone or buprenorphine exposure. *N Engl J Med.* 2010;363(24):2320-2331.
14. Volpicelli JR, Alterman AI, Hayashida M, O'Brien CP. Naltrexone in the treatment of alcohol dependence. *Arch Gen Psychiatry.* 1992;49(11):876-880.
15. Anton RF, Latham P, Voronin K, et al. Efficacy of Gabapentin for the Treatment of Alcohol Use Disorder in Patients With Alcohol Withdrawal Symptoms: A Randomized Clinical Trial. *JAMA Int Med.* 2020;180(5):728-36.
16. Krebs TS, Johansen PO. Lysergic acid diethylamide (LSD) for alcoholism: meta-analysis of randomized controlled trials. *J Psychopharmacol.* 2012;26(7):994-1002.
17. Ebbert JO, Hughes JR, West RJ, et al. Effect of varenicline on smoking cessation through smoking reduction: a randomized clinical trial. *JAMA.* 2015;313(7):687-694.
18. Walker N, Howe C, Glover M, et al. Cytisine versus nicotine for smoking cessation. *N Engl J Med.* 2014;371(25):2353-2362.
19. Courtney RJ, McRobbie H, Tutka P, et al. Effect of Cytisine vs Varenicline on Smoking Cessation: A Randomized Clinical Trial. *JAMA.* 2021;326(1):56-64.
20. Dole VP, Robinson JW, Orraca J, Towns E, Searcy P, Caine E. Methadone treatment of randomly selected criminal addicts. *N Engl J Med.* 1969;280(25):1372-1375.
21. Fudala PJ, Bridge TP, Herbert S, et al. Office-based treatment of opiate addiction with a sublingual-tablet formulation of buprenorphine and naloxone. *N Engl J Med.* 2003;349(10):949-958.
22. Rosenthal RN, Lofwall MR, Kim S, et al. Effect of Buprenorphine Implants on Illicit Opioid Use Among Abstinent Adults With Opioid Dependence Treated With Sublingual Buprenorphine: A Randomized Clinical Trial. *JAMA.* 2016;316(3):282-290.
23. Lofwall MR, Walsh SL, Nunes EV, et al. Weekly and Monthly Subcutaneous Buprenorphine Depot Formulations vs Daily Sublingual Buprenorphine With Naloxone for Treatment of Opioid Use Disorder: A Randomized Clinical Trial. *JAMA Intern Med.* 2018;178(6):764-773.
24. Ray LA, Meredith LR, Kiluk BD, Walthers J, Carroll KM, Magill M. Combined Pharmacotherapy and Cognitive Behavioral Therapy for Adults With Alcohol or Substance Use Disorders: A Systematic Review and Meta-analysis. *JAMA Netw Open.* 2020;3(6):e208279.
25. Gryczynski J, Nordeck CD, Welsh C, Mitchell SG, O'Grady KE, Schwartz RP. Preventing Hospital Readmission for Patients With Comorbid Substance Use Disorder : A Randomized Trial. *Ann Intern Med.* 2021;174(7):899-909.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

26. Orme S, Zarkin GA, Dunlap LJ, et al. Cost and Cost Savings of Navigation Services to Avoid Rehospitalization for a Comorbid Substance Use Disorder Population. *Med Care*. 2022;60(8):631-635. doi:10.1097/MLR.0000000000001743.
27. Kelly JF, Greene MC, Bergman BG, White WL, Hoepfner BB. How Many Recovery Attempts Does it Take to Successfully Resolve an Alcohol or Drug Problem? Estimates and Correlates From a National Study of Recovering U.S. Adults. *Alcohol Clin Exp Res*. 2019;43(7):1533-1544.
28. Jacobs A. Legal Use of Hallucinogenic Mushrooms Begins in Oregon. *New York Times*. Jan 3, 2023. Section D, Page 1.

Submitted 01-04-2023

VI.5 EBM in Surgical Trials- Identifying unique challenges (Anna Witkin, GSM4)

Introduction:

The gold standard for conducting clinical research, the randomized control trial, comes with some unique challenges when applied to surgical interventions. While a placebo-controlled arm is relatively simple in a medical trial, the possibility of a surgical placebo raises logistical and ethical questions. Additional factors unique to surgical trials can reduce the quality of the results such as difficulties with blinding, recruitment, and variation between individual surgeons.

However, because every surgical intervention is associated with real risk of harm to the patient, it is particularly important to ensure that there is a sufficient evidence base to support the decision to operate and guide operative strategy. Here, we walk through some of the biggest issues facing surgical research, how to assess surgical evidence for common sources of bias, and potential methods to improve evidence quality.

Skill of the surgeon:

Unlike medical trials, procedural trials can be significantly impacted by the skill of the surgeons who are participating. This can be particularly problematic if comparing two procedures, one of which is newer or less common. The surgeons would then be likely to have less experience with one procedure which might lead to worse outcomes for patients randomized to that group. Furthermore, as the trial progresses, the surgeon would have performed the newer procedure more times and likely gained competency. This could result in different outcomes for patients enrolled earlier versus later in the course of the study.

This issue can be addressed and its impact on the results of a trial limited by ensuring all participating surgeons have completed a minimum number of the newer procedure prior to performing it on a trial participant. However, this practice could limit surgeons' willingness to participate in the trial.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Placebo:

Arguably the most ethically fraught issue facing surgical trials is the use of a placebo arm. Sham surgeries are difficult to justify given the inherent risk of any operative intervention. Trial design is particularly challenging when attempting to compare operative to non-operative management, but there are some circumstances under which a surgical placebo may be ethically employed.

The fundamental document guiding medical research ethics, The Declaration of Helsinki, states that “medical research involving human subjects may only be conducted if the importance of the objective outweighs the inherent risks and burdens to the research subjects” and the “[new knowledge] can never take precedence over the rights and interest of individual research subjects”. This statement makes clear that the knowledge gained from conducting the trial is not in and of itself an adequate justification for subjecting individual participants to undue risk.

The key concept that should guide a determination of the ethics of a surgical placebo is clinical equipoise: an equal probability of benefit between two groups based on the current body of evidence. One scenario under which there may be clinical equipoise is in the case of a trial where the intervention arm is a laparoscopic procedure. In such cases the placebo may serve as a diagnostic laparoscopy and thus provide a direct diagnostic benefit to participants. Furthermore, taking part in a trial may in and of itself have an indirect benefit on participants because patients in a trial tend to do better than patients in standard care (Savulescu et al.). This may be due to the increased oversight that accompanies involvement in a trial and/or some degree of placebo effect. For these reasons, there are opportunities to ethically employ a surgical placebo arm in certain cases.

Blinding:

Due in large part to the challenges associated with a placebo control in surgical trials, it can range from difficult to impossible to keep participants and clinicians blinded as to which arm patients have been randomized. If there is no surgical placebo, both surgeons and participants will obviously be aware of whether a procedure was performed, which can be a significant source of bias. If ethically possible, a surgical placebo is the most effective way to eliminate this bias by allowing for blinding. If this is not feasible, there may be some tactics to decrease the effect of the bias such as utilizing blinded outcome evaluators post-operatively.

Enrollment:

Enrolling enough patients to adequately power a study tends to be difficult for surgical trials given that patients often have strong emotional responses to surgery, and particularly struggle to surrender control of the decision “to go under the knife”. Not only is it more challenging to enroll patients, but patients are liable to drop out or cross over if they discover they are not satisfied with the arm into which they were randomized.

While this is to some degree an inherent issue with surgical trials, optimizing the consultation process for potential participants may help ensure studies are adequately powered. The ProtecT trial (focused on prostate cancer testing and treatment) initially met with difficulty enrolling patients and focused on

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

addressing barriers to successful recruitment. Analysis of recordings of consultations along with interviews with surgeons and patients about their consultation revealed unrecognized biases were communicated by the surgeons, and that patients were often confused about the process (Blazeby et al.). After giving surgeons individual feedback and structuring consultations to maximize clear communication, enrollment rose from 30% to 65%. Providing participating surgeons with tools to make consultations clearer could produce similar results in future studies.

Conclusions:

The field of surgical research faces significant challenges which ultimately result in less robust bodies of evidence to support surgical decision-making than are warranted considering the risks of operative intervention. While many of the potential sources of bias cannot be completely eliminated and surgical trials may remain of somewhat lower quality than their medical counterparts, there are some steps that could be taken to improve the validity and generalizability of evidence.

References:

Strobel O, Büchler MW. The problem of the poor control arm in surgical randomized controlled trials. *Br J Surg*. 2013 Jan;100(2):172-3. doi: 10.1002/bjs.8998. Epub 2012 Nov 23. PMID: 23180610.

Savulescu J, Wartolowska K, Carr A. Randomised placebo-controlled trials of surgery: ethical analysis and guidelines. *J Med Ethics*. 2016 Dec;42(12):776-783. doi: 10.1136/medethics-2015-103333. Epub 2016 Oct 24. PMID: 27777269; PMCID: PMC5256399.

Lassen K, Høy A, Myrmet T. Randomised trials in surgery: the burden of evidence. *Rev Recent Clin Trials*. 2012 Aug;7(3):244-8. doi: 10.2174/157488712802281402. PMID: 22621283.

Robinson NB, Fremes S, Hameed I, Rahouma M, Weidenmann V, Demetres M, Morsi M, Soletti G, Di Franco A, Zenati MA, Raja SG, Moher D, Bakaeen F, Chikwe J, Bhatt DL, Kurlansky P, Girardi LN, Gaudino M. Characteristics of Randomized Clinical Trials in Surgery From 2008 to 2020: A Systematic Review. *JAMA Netw Open*. 2021 Jun 1;4(6):e2114494. doi: 10.1001/jamanetworkopen.2021.14494. PMID: 34190996; PMCID: PMC8246313.

Søreide K, Alderson D, Bergenfelz A, Beynon J, Connor S, Deckelbaum DL, Dejong CH, Earnshaw JJ, Kyamanywa P, Perez RO, Sakai Y, Winter DC; International Research Collaboration in Surgery (IRIS) ad-hoc working group. Strategies to improve clinical research in surgery through international collaboration. *Lancet*. 2013 Sep 28;382(9898):1140-51. doi: 10.1016/S0140-6736(13)61455-5. PMID: 24075054.

Probst P, Grummich K, Harnoss JC, Hüttner FJ, Jensen K, Braun S, Kieser M, Ulrich A, Büchler MW, Diener MK. Placebo-Controlled Trials in Surgery: A Systematic Review and Meta-Analysis. *Medicine (Baltimore)*. 2016 Apr;95(17):e3516. doi: 10.1097/MD.0000000000003516. PMID: 27124060; PMCID: PMC4998723.

McCulloch P, Feinberg J, Philippou Y, Koliass A, Kehoe S, Lancaster G, Donovan J, Petrinic T, Agha R, Pennell C. Progress in clinical research in surgery and IDEAL. *Lancet*. 2018 Jul 7;392(10141):88-94. doi: 10.1016/S0140-6736(18)30102-8. Epub 2018 Jan 18. PMID: 29361334.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Blazeby JM. Recruiting patients into randomized clinical trials in surgery. *Br J Surg*. 2012 Mar;99(3):307-8. doi: 10.1002/bjs.7818. Epub 2012 Jan 11. PMID: 22237652.

Lombardi R. Designing randomized clinical trials in surgery. *Br J Surg*. 2014 Mar;101(4):293-5. doi: 10.1002/bjs.9323. Epub 2014 Jan 29. PMID: 24477768.

Submitted 11/21/23

VI.6 Evidence-Based Psychiatry- History, Benefits, and Harms of the DSM-V (Rachel Brown, GSM4)

A Brief Introduction to the DSM and Its Relationship to EBM

The Diagnostic Statistical Manual is the basis of psychiatric diagnosis in the United States. The first edition was published in 1952, with the most recent edition, DSM-V, being released in 2013.¹ Because the DSM-III is the skeleton of the categorical diagnostic framework still used today, it is worthwhile to briefly review the cultural forces and methodological approaches taken to create the third, fourth, and fifth editions of the DSM.

DSM-III (1980)

Prior to the establishment of DSM-III, the criteria for psychiatric diagnosis was poorly standardized and lacked reliability. DSM-I and -II largely relied upon psychoanalysis as the etiologic basis for understanding psychiatric pathology. In the 1970s, up to 42% of the time, two psychiatrists offered differing diagnoses of the same patient.² The lack of diagnostic reliability led to criticism from psychopharmaceutical companies who, in the wake of the 1960s thalidomide crisis, had a heightened desire to conduct RCTs demonstrating drug safety and efficacy. The imprecision of the diagnostic labels threatened the utility of those research endeavors and drew ire from insurance companies, who sought more clearly coded criteria for the sake of reimbursements.³

At the time of the creation of the DSM-III, the term “evidence-based medicine” did not exist, and there was no formally established hierarchy for ranking evidence in terms of quality. The primary form of evidence used for the DSM-III was the expert consensus and committee voting of a 15-person Task Force led by Dr. Robert Spitzer. The procedures of the task force were explained by Spitzer: “Our general principle was that if a large enough number of clinicians felt that a diagnostic concept was important in their work then we were likely to add it as a new category. That was essentially it. It became a question of how much consensus there was to recognise and include a particular disorder”.³ While the Task Force sought to understand the existing research in the field, Spitzer noted that “there are very few disorders whose definition was the result of specific research data.”³

The advancement sought by the third edition was an improvement in inter-rater diagnostic reliability. By explicating operational criteria for mental disorders (in the form of observable descriptive features), diagnosis with improved inter-rater reliability could be completed through highly structured interviews.⁴ The consensus definitions improved basic issues like differential prevalence. In the 1950s schizophrenia was ostensibly twice as prevalent in the US as in Great Britain. However, once the DSM-III established the criterion that symptoms of schizophrenia must be present for 6 months to garner the diagnosis, American psychiatrists stopped diagnosing “acute schizophrenia” and the prevalence rates between the countries eventually equalized.⁵

DSM-IV (1994)

Similarly to DSM-III, the core focus of the fourth edition was continued improvement in diagnostic reliability. Dr. Allen Frances, who chaired the Task Force, stated, “I had no grand illusions either of seeing reality straight on or of reconstructing it whole cloth from my own pet theories. I just wanted to get the job done—i.e., produce a useful document that would make the fewest possible mistakes, and create the fewest problems for patients.”⁶ The basis of diagnosis remained the identification of a checklist of symptoms over a period of time, relying primarily upon patient recall. However, as with the prior edition, DSM-IV drew criticism for its aim of delineating diagnoses without any sort of underlying framework of etiology. Ultimately, diagnostic reliability was enhanced without any attempt at addressing diagnostic validity. Psychiatrist Paul McHugh maligned the DSM-IV for this reason, contending, “It does not speak to the nature of mental disorders or distinguish them by anything more essential than their clinical appearance. Not a gesture does it make toward the etiopathic principles of cause and mechanism that organize medical classifications.”⁷

Still, the DSM-IV improved upon the DSM-III in its methodology by incorporating principles from the then-nascent field of evidence-based medicine. In 1990, the DSM-IV held a Methods Conference to establish guidelines for Work Group members (individuals responsible for providing research summaries and recommendations to the Task Force experts) to complete literature reviews, complete with external review, prior to offering any recommendations for changes to the Task Force.⁸ Ultimately, this policy was only variably enforced.

DSM-V (2013)

In May of 2013, the DSM-V was released. Unlike DSM-IV, no procedures were set in advance to encourage or enforce the use of systematic literature reviews prior to Task Force committee meetings or the advancement of recommendations for changes to the DSM. Despite the growth of evidence-based medicine since the term was coined in 1992, the DSM-V continued to rely heavily upon expert consensus (rather than higher quality evidence) without specific measures to promote objectivity in group meetings.⁸

Benefits of DSM-V

Inter-rater reliability: If nothing else, the DSM-V offers common language and shared definitions for psychiatric diagnoses, building on Spitzer’s original intent with DSM-III.⁶ The manual offers consensus for physicians, patients, researchers, and insurance companies. It is the best tool available to clinicians to guide them in the diagnosis and treatment of patients.^{6,9}

Key to mental health service access: DSM diagnostic labels serve as an organizing principle in guiding the disbursement of mental health care and benefits. The diagnoses are “passports to insurance coverage, the keys to special educational and behavioral services in school, and the tickets to disability benefits.”¹⁰ Patients with mental illness who seek affordable medications, ADA accommodations, and Social Security benefits typically require a DSM diagnosis for access to these resources.

Harms of DSM-V

Lack of validity: The most damning criticism of the DSM-V is that its diagnostic categories lack validity. The basis of the diagnoses remains the same as earlier editions, relying upon “phenotypic features and patient recall of experience, not experimental evidence.”¹¹ Dr. Thomas Insel, former head of the National Institute of Mental Health, offered the following rebuke of the fifth edition:

The weakness is its lack of validity. Unlike our definitions of ischemic heart disease, lymphoma, or AIDS, the DSM diagnoses are based on a consensus about clusters of clinical symptoms, not any objective laboratory measure. In the rest of medicine, this would be equivalent to creating diagnostic systems based on the nature of chest pain or the quality of fever. Indeed, symptom-based diagnosis, once common in other areas of medicine, has been largely replaced in the past half century as we have understood that symptoms alone rarely indicate the best choice of treatment.⁹

Medicalization of “Normal”:

The descriptive diagnostic system outlined in DSM-V leads to binary judgment on behalf of the diagnostician about the existence or absence of mental illness. Unlike other forms of evidence-based diagnostic testing, the DSM lacks any clear-cut reference standard against which to be independently, blindly compared.¹² Instead, diagnosticians must judge behaviors as normal or abnormal, introducing value judgments based on what they view as socially acceptable behavior. One of the most notorious examples of such value judgments is the inclusion of homosexuality as a DSM diagnosis until 1973.¹² Perhaps more benign is research suggesting that normal shyness has been recast as social anxiety disorder, leading to pharmacologic treatment of distress but not disease.¹³

Ultimately, because the DSM diagnoses rely upon social construction, there is the inevitable harm that misjudgment of normal behavior leads to unnecessary diagnosis. Noting the inherent fallibility of such social constructions, Allen Frances concludes that the best definition of a mental disorder is “what clinicians treat and researchers research and educators teach and insurance companies pay for.”⁶

Barriers to Research Advancement:

Concerns stemming from the lack of validity of the DSM's diagnostic categories impact the usefulness of research using those categories. (Ironically, this concern is not dissimilar from the one about lack of reliability levied by pharmaceutical companies prior to the creation of DSM-III). DSM diagnoses are based on a categorical-polythetic framework (categorical meaning that a binary determination is made about whether the disease exists or not; polythetic meaning that a specific number and combination of symptoms, but not every possible symptom, must be present to qualify for a diagnosis).¹⁴ The consequence of the categorical-polythetic approach is considerable within-group prognostic heterogeneity (owing to differences in presentation and severity). As an example, there are 93 possible symptom combinations (from the nine symptoms comprising the major depression criteria) that would warrant a diagnosis of major depression.¹⁵ Within the realm of evidence-based medicine, the consequence is that randomization is tasked with some heavy lifting—namely, evenly distributing all the unknown confounders that stem from studying such heterogeneous diagnostic groups. Still, randomization only reliably achieves the balancing of unknown confounders over an infinite number of study replications. In reality, it is likely that at least one confounder will be unevenly distributed in any given psychiatric trial.¹⁵

Another issue stemming from the categorical-polythetic approach is the comorbidity/multimorbidity issue; most individuals who meet criteria for one psychiatric diagnosis also meet criteria for another.¹⁴ It seems that most psychiatric cases do not neatly fit into the DSM's "artificial diagnostic silos".⁵ Analyses of comorbidity and twin data have revealed shared underlying risks between different diagnostic categories. Krueger and Markon asserted that "the tendency for mental disorders to be comorbid is neither artifact nor nuisance; it is instead a predictable consequence of the involvement of common liability factors in multiple disorders."¹⁴ Such research suggests that better explanatory models may eventually supplant the DSM-V categories. Based on this line of thinking, in 2013 the head of the National Institute of Mental Health, Thomas Insel, announced a shift of research funding from the studies based in DSM-V categories to a new framework, the Research Domain Criteria (RDoC) project.⁹ His aim in the decade since has been to "begin collecting the genetic, imaging, physiologic, and cognitive data to see how all the data, not just the symptoms, cluster" so that a new psychiatric nosology may be devised based on etiology.⁹

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

A final barrier to the advancement of psychiatric research is a philosophical concern that is present in, but not to be blamed on, the DSM-V. Namely, the nebulous, experiential nature of mental illness may defy reduction to the objective measures demanded by evidence-based medicine. Numerical changes in a symptom rating scale may not fully capture the reality of what recovery or remission from mental illness looks like.¹⁵ Likewise, the span of psycho-social-cultural stressors that contribute to the manifestation of mental illnesses are likely so numerous as to defy numeration.¹⁶ Allen Frances asserts, “We will never have the perfect diagnostic system. Our classification of mental disorders will always necessarily be no more than a collection of fallible and limited constructs that seek but never find an elusive truth.”⁶ While our approximations of mental illness may improve and “precision psychiatry” may someday better capture individual nuance, the singularity of human experience is fundamentally at odds with the rigidity of diagnostic groupings and the controlled structure of experimentation.

Concluding Note: The Path Forward

Thomas Insel recently released a biography reflecting on his tenure as head of the National Institute of Mental Health over the last decade, during which he refocused the nation’s research agenda from the DSM-V to genetics and neuroscience through the RDoC project. He told the story of meeting a man whose son was dealing with schizophrenia, suicidality, and homelessness. The man told Insel, “Our house is on fire and you are talking about the chemistry of the paint. What are you doing to put out this fire?” Insel says this interaction prompted him to recognize that the billions of dollars invested in basic research have not translated to the alleviation of suffering for the nation’s mentally ill.¹⁷ Certainly the DSM-V is a fallible construct, but until the promise of a new etiology-based nosology materializes, it remains the tool available to serve patients.

In the meantime, the uncertainty afforded and flaws evinced by the DSM-V have the opportunity to be “liberating” (per Allen Frances), as a skilled clinician can add modifiers and work between diagnoses to better meet the personal needs of patients.⁶ In an op-ed for the New York Times in 2013, David Brooks similarly remarked that the skillset of a psychiatrist is rooted in their ability to navigate uncertainty:

Psychiatrists are not heroes of science. They are heroes of uncertainty, using improvisation, knowledge, and artistry to improve people's lives...The desire to be more like the hard sciences has distorted economics, education, political science, psychiatry, and other behavioral fields. It's led practitioners to claim more knowledge than they can possibly have. It's devalued a certain sort of hybrid mentality that is better suited to these realms, the mentality that has one foot in the world of science and one in the liberal arts, that involves bringing multiple vantage points to human behavior.¹⁸

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Ultimately, evidence-based psychiatry is still in its infancy. While psychiatrists have an obligation as scientists to advance the knowledge of the field, they have an equally urgent duty to alleviate the suffering of their patients through all means possible.

References

1. Blashfield, Roger K et al. "The cycle of classification: DSM-I through DSM-5." *Annual review of clinical psychology* vol. 10 (2014): 25-51. doi:10.1146/annurev-clinpsy-032813-153639.
2. Carlat, Daniel. *Unhinged: The trouble with psychiatry-a doctor's revelations about a profession in crisis*. Simon and Schuster, 2010.
3. Davies, James. "How voting and consensus created the diagnostic and statistical manual of mental disorders (DSM-III)." *Anthropology & Medicine* 24.1 (2017): 32-46.
4. Krueger, Robert F., and Kristian E. Markon. "Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology." *Annual review of clinical psychology* 2 (2006): 111.
5. Hyman, Steven E. "Diagnosing the DSM: diagnostic classification needs fundamental reform." *Cerebrum: the Dana forum on brain science*. Vol. 2011. Dana Foundation, 2011.
6. Frances, Allen. "DSM in Philosophyland: Curiouser and Curiouser." *Making the DSM-5: Concepts and Controversies*. New York, NY: Springer, 2013.v
7. McHugh, Paul R. "Psychiatry at stalemate." *Cerebrum* (2009).
8. Kendler, K. S., and M. Solomon. "Expert consensus v. evidence-based approaches in the revision of the DSM." *Psychological Medicine* 46.11 (2016): 2255-2262.
9. Insel, Thomas. "Director's blog: Transforming diagnosis." *National Institute of Mental Health* 29 (2013).
10. Satel, Sally. "Why the fuss over the DSM-5." *The New York Times* 12 (2013).
11. Levine, Robert, and Max Fink. "Why evidence-based medicine cannot be applied to psychiatry." *Psychiatric Times* 25.4 (2008): 10-10.
12. Straus, Sharon E., et al. *Evidence-based medicine E-book: How to practice and teach EBM*. Elsevier Health Sciences, 2018.
13. Lane, Christopher. *Shyness: How normal behavior became a sickness*. Yale University Press, 2008.
14. Krueger, Robert F, and Serena Bezdjian. "Enhancing research and treatment of mental disorders with dimensional concepts: toward DSM-V and ICD-11." *World psychiatry : official journal of the World Psychiatric Association (WPA)* vol. 8,1 (2009): 3-6. doi:10.1002/j.2051-5545.2009.tb00197.x
15. Gupta, Mona. "Does evidence-based medicine apply to psychiatry?." *Theoretical medicine and bioethics* vol. 28,2 (2007): 103-20. doi:10.1007/s11017-007-9029-x.
16. Tripathi, Adarsh et al. "Biopsychosocial Model in Contemporary Psychiatry: Current Validity and Future Prospects." *Indian journal of psychological medicine* vol. 41,6 582-585. 11 Nov. 2019, doi:10.4103/IJPSYM.IJPSYM_314_19.
17. Barry, Ellen. "The 'Nation's Psychiatrist' Takes Stock, With Frustration." *New York Times*, 22 Feb. 2022.
18. Brooks, David. "Heroes of Uncertainty." *New York Times*, 27 May 2013.

Submitted 1-10-2023

VI.7 The Blinded Leading the Blind – Unique Methodologic Challenges in Psychiatric Research and the Implications for Modern Psychedelic Research (Joseph Tella, GSM4)

The following Q&A-styled essay is meant to serve as a foundation to help those who wish to better understand the psychiatric literature, some of its context, and the limitations often present in the methodology of modern psychiatric research including a small focus on contemporary psychedelic trials. Many of the included subjects, such as controls and blinding, are often relatively straightforward in the research of other healthcare fields. As such, this essay also explores the relevant biases related to the quirks of these concepts as they manifest in the modern psychiatric literature.

Why do we use control groups in clinical research trials?

Control groups serve as a benchmark for comparison between the results of an experimental treatment and an alternative, which is often doing nothing (e.g. placebo) or, in cases when doing nothing would be unethical, another treatment that is commonly used in the field or is a current standard of care. In trials where the control is placebo, and the intervention outperforms the control, one can feel more confident that the results of the trial are due to the experimental treatment itself, and the findings of the trial support efficacy of the experimental treatment. There are useful considerations when thinking about active agents or placebos, but let's focus for now on how controls have impacted the field of psychiatry and its research.

What has psychiatry used for controls in the past?

In trials investigating the efficacy of a psychiatric medication, as in other healthcare research, psychiatric research has often relied on placebo or a “fake medication” such as a sugar pill or plain saline. In other words, any inert substance that should theoretically have no effect on the clinical problem being researched.

In trials investigating the efficacy of a non-pharmacologic intervention such as psychotherapy, placebo does not really fit as an option for control. To address this, psychiatric research has often relied on “waitlist controls,” where the participants allocated to the control group are placed on a waitlist where they receive no psychotherapy until a pre-determined amount of time has elapsed. In other words, if the proposed psychotherapy takes 6 weeks to perform, the control group waits 6 weeks without therapy at which point their progress is compared to the experimental group, and then they may receive the psychotherapy. This is meant to mimic the “real world” where would-be patients often have to wait for availability in local psychotherapy practices, temporarily receiving “no treatment” and allow disease processes to continue their natural course (e.g. “spontaneous” improvement in their symptoms or also deterioration).

Another common control for psychotherapy research is psychoeducation or basic support, whereby control participants may receive education about their mental health or the psychiatric condition under study, or they may be able to talk with clinicians or research staff via basic supportive conversation, but receive no actual psychotherapy. These choices may control for the sometimes psychologically therapeutic nature of receiving attention from clinicians or basic human interaction, as opposed to truly “doing nothing” as in waitlist controls.

What challenges do these historical controls pose in psychiatric research?

Generally speaking, the main problem posed by placebo controls is the potential activity of the “placebo effect” which is discussed in greater detail below.

Waitlist controls create several potential challenges. While they are generally viewed as being relatively immune to the placebo effect, and certainly reflect a likely reality for would-be patients in the real world, the act of being placed on a waitlist still has the potential to generate bias. Being placed on a waitlist may come as a disappointment to participants, which can exacerbate symptoms of depression or anxiety and thus exaggerate the comparative impact of psychotherapy. This effect might become more pronounced the longer the waitlist time is designed to be, which also introduces potential ethical questions for researchers. Also, being placed on a waitlist is different from “life as usual,” whereby soon-to-be patients who are placed on waitlists know they will eventually receive treatment, and thus may be less inclined than the average person to enact lifestyle changes which might improve their psychiatric symptoms, thus making this control less similar to “real life.” Notably, several meta-analyses have called the use of waitlist controls into question, demonstrating diminished response of waitlist participants to their eventual treatment, inferiority as compared to other control conditions, and the loss of significance for findings derived from waitlist-controlled studies after controlling for factors such as recruitment methods and length of follow-up.¹⁻³

Psychoeducation and basic support also have potential problems, but essentially by design. Because basic attention and human interaction may have benefits for certain personalities, they might cause more improvement in psychiatric symptoms than “doing nothing” would, and thus potentially underestimate the effectiveness of a psychotherapy when used as a control. As such, psychoeducation and basic support function as relatively “active controls,” meaning they may have an effect on the condition being studied, but a relatively weak one. As such, readers should deliberately account for this when interpreting the findings of studies that elect to use active controls.

What is blinding?

When we separate participants in a trial into the control group and the experimental group for the purposes of determining whether an experimental treatment is effective, in an ideal case, both participants and the research teams should be unaware of whether a given participant is in the experimental group or the control group. The concealment of group allocation is referred to as “blinding.” Trials where only the participant is unaware of their allocation are called “single blind” and

trials where both participants and researchers are unaware are called “double blind.” Trials where all parties are aware of participant allocation are referred to as “open-label.”

Why do we blind in research trials?

Blinding is important because it reduces the potential bias in the findings of a trial, which in turn increases the internal validity of the trial. Studies that are not blinded have been reliably shown to over-estimate the effect size of experimental treatments.^{4,5} This effect is believed to be (at least primarily) due to a phenomenon called “expectancy.”

What is expectancy?

Expectancy, or the “observer-expectancy effect,” is a phenomenon in which a person’s opinions, attitudes, and expectations can influence their interpretation of events. It is related to confirmation bias, which is the tendency to search for and favor information that confirms or supports a person’s prior beliefs.⁶

For example, if a participant who is suffering from a certain health problem is aware they are receiving placebo instead of new experimental treatment, they may expect that their condition will worsen because their only “treatment” is an inert sugar pill. This in turn may make them more sensitive to the symptoms of their condition, and they may report or even be observed to be doing worse than if they had been aware that they were receiving the experimental treatment. This effect is particularly prevalent in conditions that are defined by subjective interpretation, such as mental health disorders. In fact, the disappointment resulting from the knowledge of taking placebo may itself exacerbate phenomena such as rumination, which can genuinely worsen a mental health disorder such as anxiety and depression, thus impacting a participant’s condition directly and quite legitimately. On the other hand, the inherent hope that might result from the knowledge of taking a new and exciting experimental treatment may cause a participant to feel better, which could directly impact the way they interpret their symptoms, especially if those symptoms are subjective as they are in many mental health disorders such as depression. This effect would be most salient in self-report measures such as the Beck Depression Inventory (BDI) or PTSD Checklist for DSM-5 (PCL-5).

Similarly, if a researcher is aware of the allocation of a participant in a trial, there is the potential for the researcher to impart their own subjective bias into the study. For example, if a researcher who has spent a great deal of time and energy into designing a trial, applying for funding, etc. is aware that a participant is receiving the experimental treatment, they may be more inclined to interpret that participant as receiving benefit from the experimental treatment (either consciously or unconsciously). Again, this risk is even more salient in studies that use subjective measures such as clinician-administered assessments of symptomology such as the Hamilton Depression Rating Scale (HDRS) or Clinician-Administered PTSD Scale for DSM-5 (CAPS-5).

What is the placebo effect?

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

An active, experimental treatment should have a theoretical basis to impact the disease process being studied. Placebo, by definition, should be inert (e.g. plain sugar or water), and thus should have no theoretical impact on the disease process being studied. Thus, when a trial compares an experimental treatment against a placebo control group, whatever effect is measured within the control group can thus be considered the result of the “placebo effect,” a phenomenon whereby the patient reports improvement in their symptoms and disease burden despite no actual mechanism for such an improvement being evident. The mechanism of the placebo effect is likely related to expectancy, which is discussed above, although the underlying neural circuitry and behavioral associations related to placebo are active targets of investigation.²¹ There is also the related “nocebo effect” whereby a patient’s expectations of a drug may exacerbate their experience of the adverse or side effects of a drug. It is unclear whether placebo or nocebo are the result of the same or different underlying neural circuitry, but both appear to be subject to manipulation through behavioral conditioning, neuromodulation, and even some pharmacological interventions.²¹

The placebo effect is particularly relevant in psychiatric research, as the subjective nature of the symptoms of many psychiatric complaints such as depression or anxiety makes them particularly amenable to the placebo effect, which poses challenges when considering the “true” impact of an experimental psychiatric treatment. One meta-analysis investigating the impact of the placebo effect evaluated 114 studies with adequate design and both subjective and objective continuous variables.^{7,8} In studies investigating disease processes whose signs and symptoms are subjective and thus considered “definitely amenable to placebo” (e.g. depression, pain), the effect size for placebo was comparable to the effect size of active treatments (0.29, 95% CI: 0.06 to 0.52 compared to 0.24, 95% CI: 0.00 to 0.47). Notably, there was no difference between studies that used subjective or objective outcome measures.

How do these concepts relate to modern psychedelic research?

Modern psychedelic research is in a pickle when it comes to controls and blinding. Because psychedelics are, by their intrinsic nature, highly psychoactive and often produce intense emotional and psychological effects, both participants and clinicians often know when a participant has been allocated to the experimental treatment arm of a trial investigating the efficacy of a psychedelic therapy. For example, one trial investigating the efficacy of MDMA-assisted psychotherapy for PTSD demonstrated that 94% of participants who received MDMA were able to correctly guess their treatment arm of the trial, and 75% of the placebo group were able to correctly guess they received placebo,⁹ and another trial assessing the acute effects of various psychedelics demonstrated that participants were able to identify that they received placebo 97% of the time.¹⁰ This is further complicated by the fact that psychedelics have been the subject of much hype and excitement in the media and general zeitgeist, which only serves to enhance the potential effects of expectancy in psychedelic trials, especially when participants know they are receiving the experimental treatment. Thus, participants receiving active psychedelics may be inclined to report (or even experience!) greater improvement when compared to control participants, who may in turn experience less improvement (or even deterioration) if they believe they are receiving placebo, and especially if they held a belief

that the psychedelic treatment represented their best chance at relief from their mental health burden.

This has led to intriguing decisions for researchers in the psychedelic space when it comes to controls. Niacin, which can produce skin flushing and a warm sensation has been used as control in some trials^{11,12} in the hopes that control participants might conflate these effects with psycho-activity. Several trials have elected to use much lower, “non-psychoactive” doses of the experimental treatment as a control,^{13,14} effectively controlling for the underlying biological activity of the agent, thus rendering the psychedelic experience itself to be considered the therapeutic effect of the experimental treatment. And some trials have elected to use a crossover design,^{15,16} which allows participants to serve as their own controls. There are pros and cons to each of these designs, and each requires special considerations when interpreting their respective findings. Just because an individual trial will inevitably have some flaws does not make it “junk science,” but it does necessitate that readers abstain from drawing final conclusions about a new therapeutic agent until several adequately designed studies begin to demonstrate consensus. Patience remains a virtue in the psychedelic space as in the rest of life.

If expectancy and the placebo effect can actually cause symptomatic improvement, should we even care or worry about it?

Some have voiced a reasonable question: if the placebo effect and expectancy can genuinely improve a patient’s experience of their health condition, should we stop worrying about it as a confounder and simply harness it as its own clinical tool?¹⁷⁻²¹ This is a question with no easy answers. From a purely clinical standpoint, physicians and other healthcare workers have likely always applied the use of expectancy, intentionally or not, if for no reason other than to simply help provide hope for their patients who are often suffering. However, the deliberate and active use of concealed placebo as a treatment is ethically dubious at best – most of us enter medicine to be healers, not snake oil salesmen.

That being said, the efficacy of many widely prescribed treatments such as antidepressants has been very reasonably called into question, with expectancy potentially accounting for many of the perceived benefits of these medications. Others have even explored the idea of “open-label placebo” as both its own treatment as well as a strategy for extending the efficacy of true, active medications.²¹ The implications of these concepts become even more challenging to reconcile when one considers the financial stakes and the interests of pharmaceutical corporations.

So when we are considering new therapeutic agents in the pipeline, as physicians, we bear the burden to thoughtfully consider the evidence. While no trial is perfect, we must remain skeptical when we read articles, interpret their findings, and then look to apply them in a clinical decision posed to us by a person seeking our help. We must be honest with not only our patients, our peers, and the general public, but also with ourselves. This means scrutinizing the evidence, and maintaining a high threshold for the quality of the research that we both design and consume, and later rely on when we advise our patients. And in the case of clinical research spaces where certain methodologic flaws are seemingly unavoidable norms (e.g. controls and blinding in psychedelic research), we must be patient and allow

the evidence – either supportive or discouraging – to accumulate from a multitude of trials, and ideally from a variety of study designs before casting our final judgments.

References

1. Khan A, Faucett J, Lichtenberg P, Kirsch I, Brown WA. "A systematic review of comparative efficacy of treatments and controls for depression." *PLoS One*. 2012;7.
2. Barth J, Munder T, Gerger H, et al. "Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis." *PLoS Med*. 2013;10.
3. Palpacuer C, Gallet L, Drapier D, Reymann JM, Falissard B, Naudet F. "Specific and non-specific effects of psychotherapeutic interventions for depression: Results from a meta-analysis of 84 studies." *J Psychiatr Res*. 2017;87:95-104.
4. Jüni P, Altman DG, Egger M. "Assessing the quality of controlled clinical trials." *BMJ*. 2001;323:42-46.
5. Schulz KF, Chalmers I, Hayes RJ, et al. "Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials." *JAMA*. 1995;273:408–412.
6. Nickerson, Raymond S. "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of General Psychology*. 1998;2:175–220.
7. Wampold, B. E., Minami, T., Tierney, S. C., Baskin, T. W., & Bhati, K. S. "The placebo is powerful: Estimating placebo effects in medicine and psychotherapy from randomized clinical trials." *Journal of Clinical Psychology*, 2005;6:835–854. 6.
8. Wampold BE, Imel ZE, Minami T. "The placebo effect: "relatively large" and "robust" enough to survive another assault." *J Clin Psychol*. 2007;63:401-3; 405-8.
9. Mitchell JM, et al. "MDMA-assisted therapy for moderate to severe PTSD: a randomized, placebo-controlled phase 3 trial". *Nat Med*. 2023;29:2473-2480.
10. Ley L, et al. "Comparative acute effects of mescaline, lysergic acid diethylamide, and psilocybin in a randomized, double-blind, placebo-controlled cross-over study in healthy participants." *Neuropsychopharmacology*. 2023 Oct;48(11):1659-1667
11. Ross S, Bossis A, Guss J, et al. "Rapid and sustained symptom reduction following psilocybin treatment for anxiety and depression in patients with life-threatening cancer: a randomized controlled trial." *J Psychopharmacol*. 2016;30(12):1165-1180.
12. Raison CL, et al. "Single-Dose Psilocybin Treatment for Major Depressive Disorder: A Randomized Clinical Trial." *JAMA*. 2023 Sep 5;330(9):843-853.
13. Luoma JB, Chwyl C, Bathje GJ, Davis AK, Lancelotta R. "A Meta-Analysis of Placebo-Controlled Trials of Psychedelic-Assisted Therapy." *J Psychoactive Drugs*. 2020;52(4):289-299.
14. Goodwin, GM et al. "Single-Dose Psilocybin for a Treatment-Resistant Episode of Major Depression." *N Engl J Med*. 2022 Nov 3;387(18):1637-1648.
15. Becker, AM et al. "Acute Effects of Psilocybin After Escitalopram or Placebo Pretreatment in a Randomized, Double-Blind, Placebo-Controlled, Crossover Study in Healthy Subjects." *Clin Pharmacol Ther*. 2022 Apr;111(4):886-895.
16. Ross, S et al. "Rapid and sustained symptom reduction following psilocybin treatment for anxiety and depression in patients with life-threatening cancer: a randomized controlled trial." *J Psychopharmacol*. 2016 Dec;30(12):1165-1180.

17. Rommelfanger, K. "A role for placebo therapy in psychogenic movement disorders." *Nat Rev Neurol.* 2013;9:351–356.
18. Burke MJ, et al. "Leveraging the shared neurobiology of placebo effects and functional neurological disorder: a call for research." *J Neuropsychiatry Clin Neurosci.* 2019;32(1):101–104.
19. van Osch, M et al. "Specifying the effects of physician's communication on patients' outcomes: a randomised controlled trial". *Patient Educ Couns* 2017;100(8):1482–89.
20. Butler M, Jelen L, Rucker J. Expectancy in placebo-controlled trials of psychedelics: if so, so what?. *Psychopharmacology (Berl).* 2022;239(10):3047-3055.
21. Tu, Y., Zhang, L. & Kong, J. Placebo and nocebo effects: from observation to harnessing and clinical application. *Transl Psychiatry.* 2022;524(12).

Submitted 2/17/2024

VI.8 Conducting Research with Native American Communities: Barriers and Considerations (Chenin Ryan, GSM4)

Background:

Native American communities have some of the worst health disparities compared to other US racial/ethnic groups, including lower life expectancy (5.5 years less) and higher mortality from diseases including diabetes mellitus and chronic liver disease¹. Nevertheless, while disparities within these communities are rampant, research partnerships are often limited. A key aspect of research collaboration involves tribal sovereignty, recognized by the United States Supreme Court in the 1800s, in which established treaties give tribes the authority to govern and enforce laws regarding education, health, and culture². This sovereignty also extends to the governance and ownership of any data collected on a tribe and is essential to recognize when collaborating with any indigenous community.

Conducting research with Native American communities also involves acknowledging historical wrongs and an understanding that some cases of past research have led to community stigmatization and a violation of trust. For example, an epidemiologic study initiated by a state health department in the Southwest noted an outbreak of syphilis on a

reservation³. Local newspapers noted the study's findings, leading to the local non-native communities ostracizing adults and children from that reservation⁴.

In other cases, research collected from Native communities was used for purposes the tribe was unaware of. In one study investigating the genetics of diabetes among the Havasupai tribe, tribal participants were unaware that their blood samples were also used in unapproved research to examine genetic markers for schizophrenia⁵. When it was discovered that this research was used for purposes that were not agreed upon and that the participant's blood samples were distributed nationally to other researchers, the tribe's trust in the researchers who were studying the impact of diabetes on this community was irrevocably broken.

Highlighted below are actionable items one might consider when conducting research with Indigenous populations.

Important Considerations for Tribal Community Research⁶

1. Build Relationships with the Tribal Government and Spiritual Leaders

The partnership between the tribe and the researchers is crucial in designing and implementing research. Meetings should be arranged to build trust among leaders within the community. This may also involve working with a community member to act as a consultant and provide expertise. Full transparency of the nature of the research should be provided to both the leaders and members of the community out of respect and to build trust.

2. Understanding the culture of the community

Prior to any research efforts, time should be spent understanding the tribe's unique customs as well as having an awareness of any past trauma the community may have faced, especially involving research. For instance, health researchers in the past have neglected the cultural significance of tobacco products in Native communities. In a successful study

among Lakota elders, researchers acknowledged the importance of ceremonial tobacco in traditional ceremonies within this community. They generated an instrument for asking tobacco-use questions that aligned culturally with the values of the Lakota participants, highlighting the importance of addressing customs and beliefs in research⁷.

3. Awareness of Community Events

Set dates for important events such as cultural ceremonies and rituals should be considered when setting schedules and deadlines. In addition, deaths within the community may briefly halt efforts and researchers should be aware of and respect these potential delays.

4. Community Protocol and Research Results Approval

Any research protocols should be approved by the tribe's Institutional Review Board (IRB) if they have one set in place, especially regarding how data is collected. In addition, approval for how results will be published and distributed to the community is paramount. The tribe should ultimately have the final say in how the data is distributed, and researchers should be transparent about study results.

5. Consider more Unique Research Methods

Many indigenous communities prioritize passed down oral histories and traditions. Researchers should consider utilizing mixed-methods qualitative research to add cultural relevance in with data that is collected.

6. Assure Credit is Given and be Aware of Cultural Rights

Community members who provide mentorship and guidance should be given credit in any reports or publications for their assistance. This may include giving co-authorship to key community members.

7. Research Sustainability

Once time has been spent building relationships and trust with a community, efforts should be made to find ways for any data collected to be used to create meaningful impact to community members. Therefore, there should be an awareness of the limited time frame

given through grant funding, and additional means for supporting long-term sustainable community efforts should be considered.

Conclusion

While the list of above actions adapted from the Native American Center for Excellence is not exhaustive, these should be given high priority when conducting research with Indigenous populations⁶. Regardless of what a research initiative is focused on, these projects should ultimately always been seen as a true partnership between the community and the researcher, and respect for the culture and beliefs of one's study participants should always take priority.

References:

1. The Indian Health Service. Indian Health Disparities. 2019. https://www.ihs.gov/sites/newsroom/themes/responsive2017/display_objects/documents/factsheets/Disparities.pdf.
2. Rhodes KL, Echo-Hawk A, Lewis JP, V LC, D ES, D AD. Centering Data Sovereignty, Tribal Values, and Practices for Equity in American Indian and Alaska Native Public Health Systems. *Public Health Rep.* 2023;333549231199477.
3. Gerber AR, King LC, Dunleavy GJ, Novick LF. An outbreak of syphilis on an Indian reservation: descriptive epidemiology and disease-control measures. *Am J Public Health.* 1989;79(1):83-85.
4. Davis SM, Reid R. Practicing participatory research in American Indian communities. *Am J Clin Nutr.* 1999;69(4 Suppl):755S-759S.
5. Cochran PA, Marshall CA, Garcia-Downing C, et al. Indigenous ways of knowing: implications for participatory research and community. *Am J Public Health.* 2008;98(1):22-27.
6. Native American Center for Excellence (NACE). STEPS FOR CONDUCTING RESEARCH AND EVALUATION IN NATIVE COMMUNITIES.1-18. <https://www.samhsa.gov/sites/default/files/nace-steps-conducting-research-evaluation-native-communities.pdf>.
7. Manson SM, Garroutte E, Goins RT, Henderson PN. Access, relevance, and control in the research process: lessons from Indian country. *J Aging Health.* 2004;16(5 Suppl):58S-77S.

Evidence Based Medicine Study Guide
EBM Elective
Department of Medicine

Submitted 12-4-23
