

EBM Fundamentals

I. Origins and Evolution of Evidence-Based Medicine	3
II. Finding Evidence	4
Treatment Questions and Introducing PICO	4
Diagnosis Questions and Extending PICO	6
III. Staying Current with Evidence	7
Evidence-Alerting Tools	7
Using Preappraised, Structured Clinical Databases.....	8
IV. Trustworthiness of Information	8
Avoiding Untrustworthy Medical Information.....	9
When Trusted Information Should Change Your Current Practice	11
V. Study Design	12
Study Design Types.....	12
Study Designs Included in DynaMed/DynaMedex	14
Evidential Hierarchies	14
VI. Study Outcomes	16
Patient-Oriented and Disease-Oriented Outcomes.....	16
Continuous Outcome Measures.....	17
Dichotomous Outcome Measures	17
Absolute and Relative Differences in Dichotomous Measures.....	17
Interpreting Dichotomous and Continuous Measures at the Point of Care.....	18
VII. Validity	19
External Validity.....	19
Internal Validity	20
VIII. Bias	20
Bias During Study Planning.....	21
Bias During Study Execution	23

Bias During Data Analysis and Publication 26

IX. Critical Appraisal 27

 Understanding and Appraising Treatment Studies 28

 Understanding and Appraising Studies About Diagnostic Tests 32

 Understanding and Appraising Systematic Reviews and Meta-Analyses 35

X. Guidelines and Recommendations..... 41

 Where Do Clinical Practice Guidelines Come From?..... 42

 Potential Problems with Guidelines 43

 Critical Appraisal of Guidelines and Recommendations 44

 The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) System..... 45

References..... 46

EBM Glossary 48

I. Origins and Evolution of Evidence-Based Medicine

Evidence-based medicine has been evolving since the mid-1700s when the Scottish naval surgeon James Lind published his findings on treating scurvy with lemons and oranges.¹ Dr Lind is credited with designing the first clinical trial wherein groups of sailors with similar living conditions and stages of illness were given 1 of 6 different treatments, including citrus fruits. As is now common knowledge, vitamin C proved to be an effective treatment for scurvy. However, despite the publication of this evidence in 1753, it took decades for this information to be implemented into clinical practice, with many people suffering in the meantime.

The 1990s marked an important period of growth for the paradigm of evidence-based practice, with a movement to establish a better empirical basis for the practice of medicine.² The term “evidence-based medicine” (EBM) was first coined in 1991 by the Canadian physician Gordon Guyatt when he wrote: *“Evidence-based medicine uses additional strategies [rather than looking to an authority to resolve issues of patient management], including quickly tracking down publications of studies that are directly relevant to the clinical problem, critically appraising these studies, and applying the results of the best studies to the clinical problem at hand.”*³ Guyatt and several others from McMaster University partnered with academicians from the United States to form an international EBM working group in the 1990s, which pushed forward the important concept of applying evidence to clinical scenarios rather than limiting the focus to the quality of the studies. In 1996, David Sackett, an American-Canadian physician also considered to be a pioneer in the field, published an important paper on why evidence-based medicine, which he described as *“the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients,”* was not just for those in ivory towers and armchairs.⁴ It was also during this time that systematic reviews were widely recognized as the highest quality evidence, and the Cochrane Collaboration was established.

While the modern practice of evidence-based medicine still requires the skills of literature retrieval, critical appraisal, and information synthesis, the importance of considering patient values and preferences has become more and more emphasized. This expanded concept has been described as “evidence-informed decision-making.” However, as the acquisition of medical knowledge grows exponentially, we still encounter delays in adopting some evidence-based practices in this modern era not unlike those James Lind encountered in the 1700s. Factors likely contributing to this problem include the challenge of keeping up with new literature, medical training that historically relies more on “what should work” rather than “what has been shown to work,” and the potential impact of industry on information dissemination and guideline development.⁵ The chapters that

follow aim to provide the basic skills required for the modern practice of evidence-based medicine, as well as evidence for why such practice matters.

II. Finding Evidence

Practicing EBM requires the realization that some information is more likely to represent the truth than other information. However, even if someone has all the knowledge and skill required to interpret and apply the best-available evidence, finding it is a crucial first step. Finding evidence starts with asking answerable questions. Importantly, we must also focus our question on *forward* questions – what does work? – as opposed to background questions of how and why it works. In addition, there are different areas in which one might consider asking questions, such as treatment, diagnosis, prognosis, and risk factors for a condition or disease. Here, we will focus on what are perhaps the 2 most common question types: treatment and diagnosis.

Treatment Questions and Introducing PICO

Clinical questions should be focused to be practical and relevant. For instance, asking a question like, “Do statins work?” is too broad to generate an actionable answer for clinical practice. Asking “Do statins prevent cardiovascular events?” is a bit better, but still not quite specific enough. This example sets the stage for a well-known and widely used acronym first introduced by Andy Oxman that helps us standardize our clinical questions and ensure they are appropriately focused: the **PICO** format.

PICO Element	Asks the Question	Example
P atient/Population	About whom or what population are we trying to answer a question?	<i>"In patients with established cardiovascular disease ..."</i>
I ntervention/ Exposure	What is the intervention or exposure about which we have a question?	<i>"... do statins ..."</i>
C omparison	To what are we comparing the intervention/exposure of interest?	<i>"... compared to placebo or no treatment ..."</i>
O utcome	About what outcome do we want to know?	<i>"... preventing cardiovascular events?"</i>

As you can see, when modeled after PICO, the original question "Do statins work?" became much more meaningful: "In patients with established cardiovascular disease, do statins work better than placebo in preventing cardiovascular events?"

In some cases, we may even be more specific. For instance, in the above example, we set **O** to be "cardiovascular events," but this might be comprised of different events across different studies. Some studies might include having a revascularization procedure for angina or a hospital readmission in their definition for "cardiovascular events," whereas other studies might only include so-called "hard" outcomes, such as nonfatal myocardial infarction, nonfatal stroke, cardiovascular mortality, and all-cause mortality, while others yet may only focus on surrogate endpoints like cholesterol level. Whatever the case may be, working with the PICO framework helps us ensure we are asking focused questions in a standardized format, which will be critical to focus our search for evidence.

It's important to note what descriptors are most important for Patient/Population when formulating questions about treatments. Students often quickly list off age and sex, but these factors may not be as important as risk or comorbidities. Age usually doesn't need to be broken down any more specifically than infant, child, adult, or elderly person. Sex may be important only for studies related to sex-specific health issues. Remember, we are formulating questions to find evidence that has been done, and studies of say, cardiovascular events are more likely to enroll participants of both genders. What should stand out, however, is risk – for instance, high,

moderate, or low risk of a major cardiovascular event. Outcomes from a study that only evaluated adults at high risk of cardiovascular events cannot necessarily be generalizable to a low-risk population.

A similar consideration applies to **C**, the comparator. In most questions stemming from clinical practice, **C** will be standard of care, or alternative treatments, or no treatment. Placebo-controlled studies are requested by regulators, like the FDA, as they help determine if an intervention has a true effect or not **after accounting for the placebo effect**. In routine clinical practice, however, patients never receive placebo, and one would almost never phrase a PICO question using placebo as **C**. As a matter of fact, the comparator is not very useful for the search, and often we will need to rely on the comparators that were used in the studies we identify.

Diagnosis Questions and Extending PICO

The PICO strategy can be applied to questions about diagnosis as well, but this can be a little less straightforward. Consider, for instance, the question of whether B-type/brain natriuretic peptide (BNP) is helpful in making the diagnosis of heart failure. A well-formed question in the PICO format may look something like this:

PICO element	Asks the Question	Example
P atient/Population	About whom or what population are we trying to answer a question?	"In patients who present with undifferentiated respiratory symptoms ..."
I ntervention/ Exposure	What is the intervention or exposure about which we have a question?	"When compared against the gold-standard test, echocardiography"
C omparison	To what are we comparing the intervention/exposure of interest?	"... compared to placebo or no treatment ..."
O utcome	About what outcome do we want to know?	"... help discern which patients do and do not have heart failure as the primary cause of their symptoms?"

There are a few things to note here. First, the Patient/Population reflects some degree of diagnostic uncertainty: these people have a relatively undifferentiated presentation for which an underlying diagnosis needs to be established. In this example, we are looking at BNP as studied in patients with undifferentiated respiratory symptoms. When applied to diagnostic questions, the population under study has some probability of either having or not having the condition. If they clearly already have it or cannot develop the condition, they are not the right population to study. As we move down the chart, you can see that Intervention/Exposure is simply the diagnostic test of interest (BNP), and the Comparator (echocardiography) refers to the diagnostic gold standard test which is enclosed in parentheses because it is sometimes implied rather than explicitly stated, although the search for evidence is more efficient when this is explicit. Finally, you can see the Outcome is if BNP is helpful in determining whether the patient has heart failure as a component of their presenting symptoms.

III. Staying Current with Evidence

In 1950, medical literature was estimated to double every 50 years. In the year 2020, the doubling interval was estimated to be only 73 days.⁶ So, how can a practicing clinician possibly keep up? The answer is: you can't.⁷ It is quite literally impossible for any single person to stay current on all the health care literature published every day, even after limiting it to scope of practice. You have to have a system if you want to have any hope of keeping up.

There are 2 main strategies that can help you keep up with medical literature: the use of evidence-alerting tools and trustworthy structured clinical databases.

Evidence-Alerting Tools

Evidence-alerting tools are an easy way to keep up with health care literature. These tools deliver new information to you (usually via email) without you having to invest any energy to get it. DynaMed/DynaMedex offers evidence-alert features that can be tailored to your clinical interests, as well as a weekly summary of new evidence called the EBM Focus, which is available without a DynaMed/DynaMedex subscription. Additionally, many high-impact journals offer free access to tables of content and send email alerts when these and other important content (such as early online articles) are published on their websites.

Using Preappraised, Structured Clinical Databases

Trustworthy, searchable clinical resources, such as DynaMed/DynaMedex, offer preappraised structured evidence based on systematic searching of the body of health care literature. This is probably the best strategy to both stay current on evaluation and management of clinical conditions relevant to your practice and to find evidence-based answers to clinical questions at the point-of-care. Search efficiency improves with more frequent use and familiarity with whichever resource you use, so the more you look things up, the quicker you get at finding the answers you need. And just like a chef might use different knives to cut different types of foods, different clinical situations can call for different resources.

Clinical databases like DynaMed/DynaMedex are also a great resource for teaching medical students and residents the importance of keeping up with the most recent evidence and using an evidence-based approach for clinical care of patients. These clinical databases can also be used in a number of clinical settings, such as inpatient rounds and precepting of outpatient encounters.

It's important to understand different clinical resources are created in different ways. Resources can be more or less systematic in their approach to updating their content with new information and may be more or less transparent in their approach to critical appraisal and guidance. Evidence-informed decision-making requires consideration of the *best available evidence*. An ideal clinical resource will provide a user with information that incorporates and synthesizes the most current, highest-quality evidence in a transparent and easy-to-understand format and will require the least amount of work to find the information. The bottom line is that in order to keep up with health care literature, you need to know the resources available to you and use them often.

IV. Trustworthiness of Information

We inherently seek different types of information to inform our decision-making, including facts, context, and downstream effects of the decision. We may also solicit opinions and recommendations from friends and colleagues, which often reflect their values. For clinical decisions, we must weigh the relative trustworthiness of all the information gathered. Then we can make a decision.

Avoiding Untrustworthy Medical Information

Determining the trustworthiness of medical knowledge is not easy, and unfortunately there are no regulated systems in place that ensure sharing of only trustworthy information. Accurately assessing the trustworthiness of clinical information requires an understanding of the reasons information may be untrustworthy, what was done to minimize these influences, and transparency when these influences are unavoidable. The peer-review process offers some safeguard on the publication of untrustworthy information, but the process is only as good as the people involved and can be highly variable in quality.

Information is often not trustworthy because it is distorted by bias (systematic deviation) or confounding (the impact of unknown effect modifiers). Recognizing bias and potential for confounding is critical in the safe and appropriate application of information to clinical decision-making. Broadly, there are 4 major types of bias:

1. **Incomplete information** – selection bias
2. **Misinterpretation of information** – analytic bias
3. **False precision** – oversimplification bias
4. **Competing interests** – intentional bias

Let's take a closer look at these bias types:

1. Incomplete Information – Selection Bias

Selection bias is commonly considered as the risk of bias in the way participants for a study are selected, specifically with increased risk if there is lack of randomization and allocation concealment. However, selection bias extends beyond selection of participants, including risk of bias in the way information is gathered at the outset, the way missing data from participants lost to follow-up are handled, and the selection of studies to publish (publication bias is a type of selection bias). Consideration of all the available information at all steps in the research process is required to best approximate the truth and avoid misrepresentation or bias.

When evaluating and summarizing evidence reports, selection bias could also lead to ignoring clinically important results. This is common in health care because authors may emphasize results that sound most compelling, they may selectively overemphasize benefits and downplay harms, and they may focus on outcomes with the “largest” difference (mathematically or statistically) rather than the outcomes of greatest importance to patients. Recommendations and values can also be subject to selection bias. A multidisciplinary recommendations panel, ideally consisting of generalists, specialists,

methodologists, epidemiologists, and patients, is necessary on guideline committees to avoid a single “expert” unduly influencing the process, and thereby introducing selection bias.

2. Misinterpretation of Information – Analytic Bias

Much of the effort in critical appraisal involves methodological assessments of the underlying evidence. High-level data analysis requires expertise in statistics, methodology, and/or specific clinical content areas. This high-level data analysis is expected of research groups before data are published but is also part of the process of critical appraisal of evidence that needs to occur by the scientific community and as part of the peer review process. This is a very good reason why nonexperts should identify and use resources, such as DynaMed/DynaMedex, that expertly and systematically critically appraise evidence, a big part of which is assessing for analytic bias.

3. False Precision – Oversimplification Bias

It's important to remember that statistics are only a representation of the truth, not the truth itself. Even the most rigorous, perfectly designed studies will still have some degree of inescapable variation in the data, termed random error, which can affect the results of any given study or body of evidence. Throughout the health care literature, statistical significance has been misinterpreted as a yes/no concept in oversimplified conclusions. The problem is further compounded by summary reports referring to cited studies as if this yes/no dichotomy is correct. For example,

If $p < 0.05$, there is a statistically significant finding. This does not mean the finding is clinically significant. It also does not mean the finding is true.

If $p > 0.05$, there is a failure to find statistical significance. This does not mean there is an absence of effect.

4. Competing Interests – Intentional Bias

Competing interests (or conflicts of interest) can occur with any person involved in the development of knowledge representation – authors, reviewers, and editors. Competing interests may be financial, but can also be related to intellectual, philosophical, professional, or even political gain. Many of these conflicts are rooted in a misplaced desire to be on the “right” side of an argument rather than getting it right. Competing interests may be conscious or unconscious, but they involve some intent to display knowledge for reasons other than for knowledge-sharing and seeking the closest possible representation of the truth. Examples include study funding by a drug company or guideline committee

members who authored major related trials or with intellectual conflict based in long-standing clinical practice.

When Trusted Information Should Change Your Current Practice

Knowing whether to change practice based on new information can be very difficult, especially when the evidence conflicts with the opinions of experts, your own well-established habits, or the known or assumed underlying pathophysiology of a disease. There is inherent uncertainty in abandoning old practices that may seem to have worked in the past. There is no single criterion to use to determine whether practice should change based on new, trusted information. The trustworthiness of information comes on a scale rather than as a yes or no answer.

Medical research relies on probability and averages – the probability that we will cause harm or provide benefit to the average patient. However, no patient is average, and for each patient, any probability will translate into a binary condition, experiencing or not experiencing some event. Some people will benefit from treatments that, on average, have not been shown to work. Also, even for treatments that are beneficial most of the time, some patients will be harmed.

The best way to become comfortable with changing practice based on new trusted information is to be aware of the evidence that supports your current practice and its limitations. One should be on the lookout for new evidence and be able to assess the new information for its trustworthiness. For some practices, we have strong evidence that has been confirmed in multiple independent studies that demonstrates that benefits outweigh the harms. For many current practices – in all specialties – the evidence is not definitive. While we can't necessarily wait until the evidence becomes clear to make clinical decisions - - we can't tell patients to "come back in 5 years when the evidence will be better" – knowing the current evidence that supports what you do will make it easier to change if research requires a shift in practice. Knowing the limitations of the current evidence that supports what you do (or don't do) may also be very important in making optimal use of resources. It can help you choose the most convenient or cheaper option among alternatives when we cannot determine one that is best.

More broadly, practicing medicine in a way that reflects the best available evidence is also the best way to avoid overuse, underuse, and misuse of resources. All can result in harm to patients. National initiatives, such as Choosing Wisely, are systematically trying to identify examples of widespread practices that result in overuse or misuse. These initiatives give good guidance to help decide when to abandon a long-used practice not supported by sound evidence.

More information about Choosing Wisely can be found at [choosingwisely.org](https://www.choosingwisely.org).

V. Study Design

The basic types of primary research can be broadly categorized into experimental studies and observational studies.

An experimental study is one in which the investigators control or alter 1 or more variables of interest, and in health care, such variables are most often interventions. By controlling/altering the variable(s) of interest, such research design allows us to gain insight into the effects of the variable(s).

Observational studies, on the other hand, simply seek to observe. There are still variables of interest that are being studied, but the researchers do not control or alter these variables. Rather, they collect data about these variables and then analyze their findings. Types of observational studies include cohort studies, case-controlled studies, cross-sectional studies, case series, and case reports.

Study Design Types

Study designs are described below in order of highest to lowest quality of the evidence they generate.

Systematic Review and Meta-Analysis

A systematic review is a process for identifying all available research findings to answer a specific question. These differ from summary or narrative reviews because of the rigorous methods used to identify research and decide what type of research findings will be included (for example, only randomized controlled trials [RCTs]). The systematic process helps controlling for potential bias. The identified research is evaluated for quality and a conclusion is drawn from the findings. In addition to a systematic review, a meta-analysis can be performed on the data, wherein results from different studies are properly combined and statistically analyzed together.

Systematic reviews can be performed with or without meta-analysis, but a meta-analysis almost always occurs in the context of a systematic review. A network meta-analysis is a specialized type of meta-analysis used to compare multiple interventions at the same time by combining data from direct and indirect comparisons from multiple randomized trials. However, care must be taken to properly conduct and interpret results from network meta-analyses, which can be more subject to bias due to the indirect component of the

inference. Systematic reviews are discussed in more detail in the section [Understanding and Appraising Systematic Reviews and Meta-Analyses](#).

Randomized Controlled Trial

A randomized controlled trial (also called a randomized trial) is the single-study design that best limits the potential for systematic bias, and thus best approximates the truth if carried out well. Randomization describes the process of allocating participants to 1 of 2 (or more) groups in a way that creates homogenous groups with the same average likelihood of outcomes and with the same blend of known or unknown characteristics affecting the response to the exposure of interest. This creates prognostically equal groups at the study outset, before the intervention in question is applied. In other words, randomization, if performed effectively, distributes subject characteristics evenly among the groups so the only difference among groups is the intervention being studied, thereby allowing for the inference of causal associations. Randomization maximizes the chances the observed effect truly and exclusively depends on the exposure.

In a randomized trial, a group of subjects is enrolled and assigned to an experimental group (or groups) or to a control group. The experimental group receives the diagnostic or treatment intervention being studied. The control group can receive a placebo, another treatment, including “usual care,” or no treatment. Dividing of patients can be done randomly, or in a quasi-random way such as enrolling patients based on the day of the week, birthday, or other method of assignment. Studies that use quasi-randomization to divide subjects, or no randomization at all, are at higher risk of bias. Allocation concealment, which is the blinding of the group to which a participant is being randomized at the time of enrollment (discussed later in detail in section [Bias During Study Execution](#)), is another important step to minimize bias in study design. Blinding to the actual treatment associated with the group(s) the patients are randomized to is crucial to ensure groups remain balanced during the study and outcomes are assessed equally across groups.

Cohort Study

A cohort study evaluates a group (“cohort”) of people followed over time to determine links between 1 or more characteristic(s) and 1 (or more) outcome(s). Prospective cohort studies enroll a group of exposed and unexposed patients and follow them forward in time to assess the effect of exposure (to a treatment or a risk factor) on outcomes. Retrospective cohort studies assemble a group of patients with previously collected data on both exposures and outcomes. Since these studies are observational with no variables under investigator control, they are generally considered to provide moderate-quality evidence.

Case Series

A case series is a report on the experience of treating or managing several patients (a “series”). An example of a case series would be a report on the effectiveness of a new surgical technique. These studies are at risk for a number of biases that can affect the results and are considered to be low-quality evidence.

Case Report

A case report gives details of a disease presentation, adverse event, or treatment outcome in up to 3 patients. Case reports typically describe novel or new occurrences. Case reports are considered very low-quality evidence.

Study Designs Included in DynaMed/DynaMedex

We don’t always have the ideal study type to evaluate every aspect of medical practice. For many areas of medicine, we have studies using designs considered less than ideal, but these may at times represent the best-available evidence. Accordingly, these studies are included in DynaMed/DynaMedex when better research does not exist, and they are clearly marked as either level 2 (midlevel) or level 3 (lacking direct evidence). But what does an “ideal study type” mean? This brings us to considering evidential hierarchies.

Evidential Hierarchies

In the section [Study Design](#), we introduced different study types commonly seen in health care literature. Many are likely familiar with the traditional evidence hierarchy depicted by a pyramid:



Evidence hierarchy pyramid

A fundamental issue with the pyramid hierarchy is that it is based on the assumption that the question of interest is always a cause-and-effect question. It also assumes equally robust research at each level. However, an RCT trial with significant threats to validity may not always represent higher quality evidence than a well-conducted observational study. Likewise, a suboptimally conducted systematic review and meta-analysis of RCTs may not be better than individual RCTs, especially in the case of so-called “mega-trials.” (See the section [Understanding and Appraising Systematic Reviews and Meta-Analyses](#) for further discussion.)

Partly in recognition of these issues, the Centre for Evidence-Based Medicine dropped the pyramidal scheme altogether, instead adopting a tabulated description of hierarchical evidence structures that depend on the question being asked. Along the same lines, Murad and colleagues have suggested the pyramidal structure itself is misleading and have suggested instead that we adopt a perspective where systematic reviews and meta-analyses are a lens through which we view the available evidence.⁸

Whatever approach one might use to determine the quality and hierarchy of evidence, it is critical to consider the limitations and threats to validity of whatever form of evidence is being appraised for the question at hand. This is where critical appraisal comes in.

More information about the Centre for Evidence-Based Medicine can be found at cebm.ox.ac.uk. A PDF of the hierarchical evidence table is available for download from cebm.ox.ac.uk/resources/levels-of-evidence/ocebmllevels-of-evidence.

VI. Study Outcomes

Patient-Oriented and Disease-Oriented Outcomes

Perhaps one of the most important questions one can ask about a given study is: would the findings of this study be relevant to patients? This speaks to the key consideration of patient-oriented vs disease-oriented outcomes. Patient-oriented evidence concerns things patients actually care about (in other words, things that affect if they live longer or better, such as death or visual loss as a complication of diabetes). Therefore, the outcomes studied in patient-oriented evidence are referred to as patient-oriented outcomes. They may also be referred to as clinical outcomes.

Disease-oriented evidence, also called surrogate or nonclinical evidence, addresses factors that do not have a direct impact on patients (for example, cholesterol or HbA1c levels, and imaging findings), but that receive attention due to their perceived or hoped-for association with things patients do truly care about. Therefore, disease-oriented evidence is considered less externally valid or generalizable than patient-oriented outcomes.

The main problem with using disease-oriented evidence for treatment decisions is that it doesn't always result in the presumed morbidity, mortality, or loss of function with which they were thought to be associated. For instance, a patient with diabetes could have diabetic retinopathy (microvascular disease of the eye) that never affects visual acuity. However, many patients with diabetic retinopathy receive expensive, invasive treatments intended to preserve vision that might never be lost in the first place. Making treatment decisions based on disease-oriented evidence alone risks unnecessary intervention, cost, and unintended harms for the possibility of no benefit.

The cases of prophylactic use of class I antidysrhythmics postmyocardial infarction and the use of rosiglitazone in type II diabetes are examples of how treatment decisions based on surrogate endpoints not only failed to help patients, but actually harmed them. In these cases, each intervention improved surrogate endpoints (such as HbA1c level) but also caused patients substantial harm, including increased risk of myocardial infarction with rosiglitazone⁹ and increased mortality with class I antidysrhythmics.¹⁰ Even if improving a

surrogate endpoint causes no harm, there would be no good reason to use the intervention if there is no benefit and only the possibility of unnecessary cost, inconvenience, and treatment burden.

It should be stated, however, that use of surrogate endpoints is often essential in the early phases of the research process before larger, long-term studies are conducted to investigate patient-oriented outcomes. Sometimes, it is the best evidence we have. However, at the point of care, surrogate endpoints are far less useful than patient-oriented evidence and are therefore automatically assigned a lower level of evidence.

For both clinical and surrogate endpoints, there are 2 broad classes of outcome measures reported in clinical studies: continuous and dichotomous.

Continuous Outcome Measures

Continuous outcomes are measurements with values that lie on a continuum. Examples include body weight, BP, duration of hospitalization, or pain measured on a 0- to 100-mm visual analog scale. These data are generally expressed by a measure of central tendency, such as the mean or median, for each study group. Efficacy of an intervention is usually expressed as the arithmetic difference in the average values between groups.

Dichotomous Outcome Measures

Dichotomous outcomes are categorical measures with only 2 possible states – patients either have an outcome event or they don't. An event could be a negative outcome, such as death or stroke, or a positive outcome, such as cancer remission, reduction in body weight by more than 10%, or migraine-free at 2 hours. Dichotomous outcomes are calculated as the proportion of patients in a study group that have the event, and provide a risk estimate (the probability of having the event). Differences in dichotomous outcomes between study groups can be described in either absolute or relative terms.

Absolute and Relative Differences in Dichotomous Measures

The absolute risk difference is the arithmetic difference between risk estimates and may be referred to as either “absolute risk reduction” or “absolute risk increase,” depending upon the direction of the effect. Absolute risk difference and its reciprocal, the “number needed to treat” or NNT, are simple statistics used when making point-of-care decisions. The NNT, usually rounded to the nearest integer, indicates the number of patients that would have to receive a treatment for 1 additional patient to achieve a positive outcome. (When a

treatment is associated with an increased risk of an adverse outcome, this value is referred to as “number needed to harm” or NNH.) Suppose a Randomized Controlled Trial (RCT) compares a new cardiovascular drug to a placebo, and it reports the absolute 5-year risk of death is 6% with the drug and 8% with the placebo. This corresponds to an absolute risk reduction of 2%, indicating the drug would prevent 2 deaths for every 100 patients treated for 5 years. The NNT for the drug would be 50 – for every 50 patients given the drug, there would be 1 fewer death in the group that was treated. This might or might not seem like a strong enough effect to warrant giving the drug to a specific patient, based on their individual risk profile and potential adverse drug effects.

Relative risk is expressed as the ratio of the risks between the 2 groups. When relative risk is presented without the context of absolute risk, it may be misleading at the point of care, because relative risk can amplify the apparent size of treatment effects. In the example above, the risk ratio (RR) is 6% divided by 8% or 0.75, suggesting that the probability of death over 5 years is 25% lower (in relative terms) with the drug. Published abstracts frequently present results in this way without further explanation. The problem with this kind of reporting is that the RR alone may not indicate the real-world clinical impact of that 25% reduction. The value of the RR would be the same if the absolute risks were 60% and 80% or 0.6% and 0.8%, but the absolute rates, and the associated absolute risk reductions and NNTs, would be very different. An absolute difference of 20% gives an NNT of 5, while an absolute difference of 0.2% gives an NNT of 500.

Interpreting Dichotomous and Continuous Measures at the Point of Care

Both dichotomous and continuous/discrete outcomes can provide important data for assessing the efficacy and safety of treatment options, but dichotomous outcomes are often easier for a clinician to use when making patient care decisions. For example, suppose a randomized controlled trial (RCT) compares a new pain drug to a placebo in adults with chronic back pain, and assesses the change from baseline to 1-month follow-up using a 100-mm visual analog scale (VAS). On a VAS, the patient rates pain severity by marking a location on a line 100 mm long, with 0 indicating no pain and 100 mm indicating “worst pain imaginable.” In this context, a 15-mm difference may be considered clinically meaningful.

The researchers could assess the drug’s efficacy based on either continuous and dichotomous measures: they could compare the mean pain reduction in each group (continuous) or they could compare the proportions of patients achieving threshold pain reduction of at least 15 mm (dichotomous). Let’s assume the mean pain reduction with the new drug was 20 mm compared to 8 mm with the placebo, and that this difference was

statistically significant. While the drug group achieved clinically important improvement on average, the mean difference between the drug and placebo groups was only 12 mm. Does this difference support the use of the drug? It's hard to say. Differences in continuous measures can tell us how a treatment will work on average, but since the effect will vary from patient to patient, it cannot tell us what to expect for a specific individual.

Unlike the continuous measure, differences in the dichotomous measure provide information about any patient's probability of achieving clinically important pain reduction. Let's assume the rates of at least 15-mm pain reduction are 45% in the drug group and 20% in the placebo group. The difference in these proportions gives an absolute risk reduction of 25%, indicating that an extra 25 out of every 100 patients taking the drug will have clinically important pain reduction, corresponding to an NNT of 4. Use of the absolute risk reduction or NNT may be easier for patients to understand than other ways of presenting the data and may simplify the treatment decision.

VII. Validity

There are many considerations that go into appraising the quality and certainty of evidence, each of which ultimately falls into 1 of 2 categories of validity: external validity and internal validity. In a nutshell, when appraising any given study or body of evidence, 2 questions should be asked: 1) How certain can I be that this evidence represents the truth about a given question? and 2) Can I apply this evidence to the patient or population at hand? These 2 questions respectively represent external and internal validity.

External Validity

We will start with external validity because it's the more straightforward of the 2. This approach can be valid even for use of evidence in clinical practice: if the study would not apply to the patient in front of me (if the patient would not have been eligible for enrollment in the study), why bother appraising whether or not the study is good? Another term for external validity is generalizability. This concept gauges how confident we are that the evidence in question applies to other patient populations, or more specifically, to the individual patient we are trying to help at the present moment. For instance, if there is a body of evidence for a certain intervention, but the population studied was exclusively people who were 65 years of age or older at high risk of cardiovascular events, we cannot assume the same evidence applies to people who are younger at low risk. The same goes for applying outcomes studied in a population of exclusively women who are younger to a

group of men who are older: you can't assume the same evidence applies. Although it would obviously be ideal to have robust evidence specific for all individual patients and patient populations we might encounter, we often do not enjoy such luxury. Therefore, we must make judgments about the generalizability/external validity of evidence based on the composition of the patient population studied compared to the patient or patient population we are interested in, and variables that may impact how well the evidence translates to the patient or patient population of interest. Importantly, such judgment is not simply a "yes/no" assessment. Rather, this judgment exists on a spectrum, with an upper limit of extreme confidence and a lower limit of extreme doubt. We often find ourselves between these 2 extremes.

Internal Validity

Internal validity refers to whether the methods used by the research group yield trustworthy evidence. Did the researchers study what they said they would study? This apparently simple question has a complex answer.

Research studies are designed with the intent to answer a particular question, to come as close as they can to uncovering "the truth." However, it is important to remember statistical analysis of experimental data only represents the truth and cannot be interpreted as the truth itself. Even the most rigorous, perfectly designed studies will still have some degree of inescapable variation in the data (termed random error), which can affect the results of any given study or body of evidence. Well-designed studies that do a good job of limiting threats to validity allow for a closer approximation of the truth and therefore are considered more trustworthy or internally valid. This is why carefully appraising evidence for threats to validity at all stages of the research process is imperative as we consider how much we can trust study outcomes.

VIII. Bias

Figuring out whether to trust information and use it in your clinical practice depends in large part on how valid the information is thought to be. We can apply the same concept to clinical resources – different resources are not equally valid. The word "validity," when used to describe studies, has an inverse relationship to bias. Threats to validity represent bias and can occur at all stages of a clinical study, including before the study begins, during study implementation, and after the study ends.

Bias During Study Planning

Systematic Bias and Study Design

Systematic bias occurs when study design inherently favors one outcome over another. This type of bias is present before the study even begins. This is why a well-done systematic review and meta-analysis of all well-done RCTs is often thought to be the highest quality or most trustworthy evidence. This type of study combines the benefits of randomization with the “law of large numbers,” which is a statistical theorem stating that by considering a larger number of observations, random error is reduced, and the results will be closer to the underlying “truth.” For the same reason, large multicentric mega-trials may accomplish the same result. Randomization is key to set up prognostically equivalent groups, allowing us to infer that any differences between the groups in the outcomes being studied are actually due to the intervention rather than random chance (confounding).

Allocation concealment, which prevents knowledge of the allocation sequence at the time of enrollment further prevents selective enrollment of subjects (selection bias). For example, imagine a trial assessing the efficacy of surfactant on the outcomes of premature infants. Knowledge of the pathophysiology of a therapy often leads to an assumption that the study intervention will be better than placebo and a natural preference to assign patients to the treatment rather than the placebo group. Knowledge of the study group to which an infant would be allocated could potentially affect the decision whether to enroll the child in the study or not.

Setting up prognostically equal groups through randomization allows inferences about cause-and-effect relationships and is a key reason why observational studies are considered a much weaker form of evidence. Observational studies carry a much larger risk that any observed differences in the outcome(s) of interest might be partially or even fully due to variables other than the intervention being studied. Cohort studies, case-control studies, case series, and case reports are all examples of observational studies (discussed in the section [Study Design](#)). Usually, the analysis of observational studies takes into account factors known to affect the outcome (the so-called known confounders, such as age, sex, severity of the disease, concomitant treatments, etc.) by using various forms of multivariable (or adjusted) analyses. However, one cannot adjust for unknown confounders, which can only be expected to be equalized between the treatment groups by the randomization process.

Selection of Outcomes

The selection of outcomes to be evaluated in a study occurs before a study begins and has a critical impact on validity. The use of surrogate endpoints, composite outcomes, and modeling can all introduce bias in important ways.

Composite outcomes are often used in trials to increase event rates, which are needed to demonstrate a statistically significant effect of an intervention. This involves combining multiple outcomes into 1 primary composite outcome. An example of a commonly used composite outcome is major adverse cardiovascular events, which could include heart failure, nonfatal myocardial infarction, need for percutaneous coronary intervention or coronary artery bypass grafting, or rehospitalization for cardiovascular-related illness. Often, significance of a composite outcome can be demonstrated in a shorter period of time and with fewer participants than with single-outcome measures. While there are circumstances under which use of composite outcomes makes sense or is necessary, there are a few critical risks with using composite outcomes, including switching outcomes, inflated composite outcomes, and a dominant surrogate outcome.

Switching of Outcomes

The components of the composite outcome should be determined before the study begins. Look out for situations where the original stated outcome is not reported or is downplayed, but instead a secondary (recombined) outcome is highlighted. For example, in a study on a diabetes medication, 6 efficacy outcomes were set to be combined as the composite outcome, but no significant difference was found. However, when only 3 of the 6 were combined, a significant difference was found. Switching of outcomes is a fishing expedition and a red flag that limits internal validity. The registration of trial protocols (including all their modifications) in public registries (such as clinicaltrials.gov) allows for the comparison between planned and reported outcomes, and should be checked any time the reported outcome does not make sense or there is suspicion of outcome manipulation.

Inflated Composite Outcomes

Sometimes the composite outcome is statistically significant, but none of the differences in the individual component outcomes is clinically important or statistically significant. Though possible just for sample size reasons, the results should be interpreted with caution. This can be of particular concern if the data for the components are not reported.

Dominant Surrogate Endpoint

If the composite outcome is significant but the only individual outcome that is significant is a surrogate endpoint, this is a threat to internal validity. A common example is

microvascular disease of the eye or kidney when evaluating outcomes related to diabetes. There might be no difference in stroke, myocardial infarction, death, need for dialysis, or blindness, but a significant difference in diabetic retinopathy alone (that may never lead to visual impairment) can make the composite outcome significant. In general, composite outcomes should be comprised of outcomes that are all of importance to the patient and should not mix hard outcomes and surrogate endpoints.

The use of modeling tools or simulation techniques carries some inherent risk of data misrepresentation. These methods use patterns observed in previous data to make predictions about outcomes under different conditions. Modeling is most often used either to estimate treatment effects when randomization is not possible or to model rare events, which can be done within a randomized trial. An example of this is the use of modeling in research evaluating interventions to prevent outcomes with low-event rates, such as cancer.

Another consideration is that the use of simulation or modeling tools assumes the previous evidence upon which the simulation is based is trustworthy. There are circumstances under which simulation-derived data represent the best available evidence, but these limitations must be considered when assessing the certainty with which you can trust the outcomes of simulation-based studies.

Bias During Study Execution

While experimental studies (randomized trials and systematic reviews of randomized trials) have, in theory, less bias due to study design compared to observational studies, once a trial is underway, there are many potential opportunities for introducing bias. This can be due to the way the study is carried out or how data are collected, which ultimately can limit the internal validity of the study results. Bias that occurs during study execution results in what is broadly termed, “performance bias,” whereby there are systematic differences between groups in study outcomes that have to do with events that occur after randomization. That is in contrast to “selection bias” that typically occurs prior to, or during, randomization.

We describe some common threats to internal validity during study implementation below.

Lack of or incomplete blinding of participants and/or outcome assessors

Blinding occurs when the intervention associated with the study groups to which a participant is assigned (for example, whether they are receiving the treatment or control intervention) is not known during the study. Blinding of the participant prevents the participant’s knowledge of the treatment affecting behaviors or attitudes that might

influence outcomes (performance bias). Blinding of the study investigator and health care personnel prevents the investigator from acting in a way that might affect the outcomes (observation bias). “Double blinding” typically refers to blinding of both the participant and the investigator. There can actually be up to 7 layers of blinding.

Another typical layer is blinding of the outcome assessor, when different from the investigator or study personnel. This is especially important when blinding of study personnel may not be possible, such as comparing a surgical intervention to a medical treatment. Importantly, masking of the intervention (or any of its effects) is critical to preserving the benefits of blinding. For example, if one treatment is a pill and one is an injection, ideally there would be 2 placebos such that each participant would be given 1 pill and 1 injection. Similarly, if the treatment affects a lab parameter (like INR), you would need to use sham INR monitoring in the control group to maintain blinding.

Lack of appropriate control

There are a number of different types of controls, including placebo, usual care, wait list controls, and sham surgeries, among others. Comparing a new treatment to no treatment is not the same as comparing it to a placebo. The placebo effect is a real effect, and a drug or other intervention compared to no treatment would make the effect of the drug or other intervention appear better than it actually is because it isn't tempered by placebo effect. Something is better than nothing with respect to study design.

For patients with knee pain, seeing an incision can make them feel like they “got something,” which has a placebo effect, thus the use of sham surgery as a comparator may lead to more robust study results rather than no intervention. For many investigations for mental health interventions, hope is a treatment. Being on a waiting list offers hope for some people and is a better comparator to an investigational treatment than no treatment at all.

Small sample size ($n < 30$ per group)

Going back to the law of large numbers, a small number of observations brings in a significant chance for random error, which can in turn bring you further from what we are considering “the truth.”

High loss to follow-up (typically $> 20\%$)

Also called attrition bias, a high or differential loss to follow-up is a type of selection bias wherein participants differentially withdraw or deviate from protocol between groups, which results in systematic differences between characteristics of the groups and affects data analysis. Participants who do not complete a study are statistically more likely to be sicker or more likely to die. While more than 20% is often used as a rough cut-off for a “high” loss to follow-up, what really matters is a rate of follow-up that is close enough to the

absolute difference (of interest or observed) to pose a threat. Performing an intention-to-treat analysis (rather than a per-protocol analysis) can be a solution to try to account for high or differential loss to follow-up, provided a valid form of missing outcome data imputation is performed.

Detection bias

Detection bias is a type of bias that can occur wherein perceived knowledge of treatment leads to systematic differences between groups in how outcomes are measured and verified. In diagnostic studies, it can also take the form of differential assessment when different tests are performed depending on previous test results.

Early termination for benefit

Ending a trial early is statistically more likely to overestimate the benefit and underestimate the harms of an intervention. This is because trials stopped early will naturally have lower event rates than would have been the case if the trial was completed, thus inflating the influence of random variability on the comparative results. When this leads to a significant benefit for the intervention, the trial may be stopped over ethical concerns about continuing some patients on placebo. If the random variability of a small result indicates no benefit at that time, the trial will continue and may wind up showing benefit in the larger sample. Thus, there are more opportunities to stop “when ahead,” while lack of benefit is usually only determined at the end of the completed trial.

Use of nonstandardized intervention under investigation

Using nonstandardized interventions (different treatment modalities within the same group) is like comparing apples to oranges and makes data interpretation difficult and less reliable. For example, if you were comparing exercise to no treatment, but “exercise” could be home-based stretching, physical therapy 2 times a week for 1 hour, or walking for 30 minutes a day, then the internal validity of the effect of exercise would be affected. It may be unclear whether any benefit shown was due to exercise in general or to a particular form of exercise.

Inadequate statistical power to detect differences

Power calculations that occur during study design rely on previous estimates of event rates (among other things) to determine what sample size is needed to achieve adequate statistical power to detect a difference between the intervention and the control. A study can be underpowered if study enrollment is lower than expected, in which case it is easy to identify that a study did not achieve the anticipated enrollment and power may be a threat to internal validity. However, a study could also be underpowered if event rates are lower than anticipated despite enrolling an adequate number of participants. In either case,

being underpowered puts a study at risk of type II error, or the risk of not detecting a difference that truly exists between groups (false negative).

Bias During Data Analysis and Publication

A well-designed trial carried out perfectly can still end up being untrustworthy if the data are not properly analyzed, interpreted, and published. While the principles of high-quality data analysis are beyond the scope of this tutorial, several common posttrial threats to validity include selective reporting, citation bias, confounding, inappropriate accounting for missing data, and post hoc analyses.

Selective Reporting

This is a form of bias due to failure to report all outcomes specified in the protocol or to outcome switching or changing the primary outcome during the trial to inflate the performance of an intervention. The most common form is the reporting of outcomes showing significant benefit with a treatment while omitting outcomes showing no effect or potentially showing harm. Excluding the results for the individual components of a composite outcome is a form of selective reporting.

Citation Bias

This is a form of intellectual bias resulting from the tendency to cite studies that uphold your point of view or beliefs. Citation bias can affect the validity of narrative reviews that do not use a systematic approach to identify relevant evidence.

Confounding

This is when an effect is attributed to an intervention, when in fact there is a nonintervention-related factor that confounds the results. This problem is more common with observational studies, and the most common problem is lack of consideration for the possibility of residual confounding from unknown confounders. Also, relying on unadjusted or overadjusted analyses may lead to confounded results.

Inappropriate Analysis of Missing Data

The effect of attrition bias (incomplete or differential loss to follow up) is unpredictable and difficult to adjust for. The ideal situation is to avoid missing data as much as possible. When data are missing, however, both per-protocol and intention-to-treat analyses should be performed and compared. For intention-to-treat to be valid, the missing data must be handled by one or more strategies, such as missing outcome imputation or scenario analyses. Of note, the effect of missing data can be for or against the intervention and

varies widely and often in opposite directions for efficacy and safety, or for superiority and noninferiority trials.

Here is an example: you are comparing surgery to medication to prevent death in a superiority trial. Because you randomized the groups and concealed allocation, the 2 groups are prognostically equal at the outset, and statistically 2 people from each group of 10 are going to die regardless of the treatment. Now, suppose the 2 people in the surgery group die before surgery and they are removed from data analysis because they didn't get the assigned treatment (surgery). Two people also die in the medication group, but they had already taken some of the medication and therefore are considered to have completed their study protocol and so are included in analysis. Let's say no one else dies. If these data are analyzed by per protocol analysis, it looks like 2/10 people in the medication group died, but 0/8 people in the surgery group died, and thus surgery appears to be a superior treatment. However, with an intention-to-treat analysis, both groups have equal deaths and there is no difference in treatment efficacy. This is just 1 example of how handling of 'missing' data can greatly impact data interpretation and results.

Post Hoc Analysis

This is a statistical analysis that is not prespecified and therefore determined after data is already known is fraught with bias. Post hoc analyses can be hypothesis-generating but should not be accepted as evidence yet.

In the following sections, we will show how the principles of clinical epidemiology thus far discussed are pulled together and applied to critically appraise certain types of evidence.

IX. Critical Appraisal

Systematically evaluating information for its trustworthiness is termed "critical appraisal." Critical appraisal can be applied to reports of trials, systematic reviews, and guidelines. This process can be labor-intensive with competencies in methodological, statistical, and clinical practice expertise all required to fully comprehend the trustworthiness of any knowledge component. Using clinical resources like DynaMed/DynaMedex that critically appraise the information presented is important for practicing clinicians who lack the time or expertise to critically appraise the literature themselves. However, having the basic skills to critically appraise treatment studies, diagnostic studies, systematic reviews, and meta-analyses will go a long way towards making sound evidence-informed medical decisions.

Understanding and Appraising Treatment Studies

Imagine we wanted to determine whether a medicine is effective at decreasing migraine occurrence. There are several approaches to study this question. Investigators could study its effect on the neurodynamics of rat brains to understand its pharmacologic effects. If the medicine is already in use to treat other medical problems, researchers could identify patients taking it for these other reasons and determine whether they are less likely to report migraines than similar people who don't take it. Another approach is to give the medicine to a number of study subjects and track whether they experience fewer migraines after starting it. Still another approach is to identify patients who experience migraines, randomly assign them to receive either the medicine or a placebo (or to receive either the medicine or another medicine that's already used to treat migraines), and follow them over time to see if there is a difference in reports of migraine attacks between the 2 groups.

As discussed in the section [Study Design](#), this last method describes an RCT, which is well-suited to answer questions about the efficacy of a treatment. Additionally, as discussed in the section [Validity](#), migraine occurrence is definitely an outcome that is important to the patient. Therefore, of the options discussed here, an RCT would have the best chance of giving us trustworthy information (assuming, of course, the trial was well-planned, well-executed, and well-analyzed as discussed in the section [Bias](#)).

Indeed, when seeking to investigate treatment effect (which is the causal relationship between exposure to a drug and the effect on the outcome of interest), the most methodologically robust primary study design is the RCT. In the realm of treatment, we are interested in learning about the extent to which an intervention might affect a given outcome, including the possibility the effect is either negligible or null. For instance, in the example from the section [Treatment Questions and Introducing PICO](#), we are interested in learning what effect (beyond placebo) taking a statin has on cardiovascular events in patients who have established cardiovascular disease.

However, for certain outcomes, an RCT may not be practical or possible. For instance, there will never be an RCT testing the effect of biological drugs for rheumatoid arthritis on cancer risk because it would be unethical and unfeasible to randomize people to take or not take a treatment for many years just to see if they will develop cancer. Similarly, you cannot randomize patients to have or not have a genetic polymorphism that predisposes them to a disease or that drives the response to treatment (like a hormone-responsive breast cancer). Additionally, for evaluating the risk of rare outcomes, such as unusual side effects,

practical limitations on the size of the trial needed to achieve adequate statistical power can preclude the use of an RCT to answer the question.

In other cases, we may simply not have any RCT evidence. In such instances, we may have to look to lower levels of evidence (for example, observational studies) for insights to inform practice, because these data may ultimately represent the best available evidence on the matter. However, whenever such a situation exists, we must still carefully consider the inherent weaknesses in such study designs compared to an RCT. But why is an RCT considered so good for investigating cause-and-effect relationships in the first place?

Perhaps the biggest advantage an RCT has over other primary study designs is the ability to produce prognostically equal groups (aside from the intervention being tested), including the balancing of unknown confounders. Knowing the groups are prognostically equal helps us to be able to infer a causal relationship between the outcome and the intervention being studied, if indeed a difference is found.

Let's consider a 2-armed RCT (an RCT with 2 study groups) that seeks to investigate a new type of inhaled therapy vs a placebo inhalation in reducing exacerbations in patients with cystic fibrosis. Participants are randomized to receive either the new inhaled therapy (treatment group) or a placebo inhalation (placebo group). Everyone in the trial is allowed to use other therapies for their cystic fibrosis that they normally would ("background therapies" or "usual care"). If randomization is effective, the 2 groups should have a highly comparable makeup of baseline characteristics, because any given participant has an equal chance of being in either the treatment group or the placebo group.

Unfortunately, randomization does not offer a universal guarantee of this baseline comparability between the groups. For instance, important baseline differences can occur more often in trials with smaller sample sizes (and this might even occur in our hypothetical trial of patients with cystic fibrosis given the smaller sample sizes seen in many such trials). Therefore, it is good practice to verify whether there are any important differences in the baseline characteristics of the groups in an RCT.

Additionally, differences in prognostically relevant variables can also arise after randomization. Ideally, this should be handled and reported by the researchers conducting the RCT, but this does not occur invariably. For instance, in our hypothetical trial, patients randomized to the placebo might end up using a larger dose and/or a greater number of background therapies for cystic fibrosis. To the extent this reduces the rate of exacerbations in the placebo group, the apparent effect of the new inhaled therapy could be reduced, because the use of background therapies is no longer comparable between the 2 groups (it is higher in the placebo group than in the new inhaled therapy group). This situation is very common in RCTs and can be accounted for by recording and analyzing the

use of other (nonstudy) medications as one of the study outcomes, often to ensure the new intervention does not affect the primary outcome (for example, exacerbation) but reduces the use of other (maybe more inconvenient, or toxic) medications. Examples would include noting the need for pain medication or in other cases, the need for use of corticosteroids.

The importance of prognostically equal groups in making inferences about cause-and-effect relationships (such as the effect of a treatment) cannot be overstated. This is a key reason why observational studies are considered a much weaker form of evidence for such inferences. In observational studies, there is a much larger risk that observed differences in the outcome(s) of interest might be partially or even fully due to variables, known or unknown, other than the variable of interest. Most observational studies are ineligible for a level 1 evidence rating in DynaMed/DynaMedex and are usually assigned a low certainty rating according to the GRADE methodology, which is commonly used in evaluating the quality of evidence for clinical practice guideline development .

As discussed in the section [Internal Validity](#), there are many other things to consider when assessing the internal validity of a study investigating a treatment. Things to consider include whether and how those involved were blinded, whether and how allocation concealment was ensured, whether the analysis was according to the intention-to-treat principle or a per protocol analysis, and how missing data were handled, to name a few.

Example Randomized Trial Evidence Summary from DynaMed/DynaMedex

STUDY SUMMARY

addition of aspirin to esomeprazole might reduce risk of high-grade dysplasia in adults with Barrett esophagus

RANDOMIZED TRIAL: [Lancet 2018 Aug 4;392\(10145\):400](#) | [Full Text](#)

Details ^

- based on randomized trial without blinding
- 2,557 adults (80% men) with Barrett esophagus \geq 1 cm were randomized in 2-by-2 factorial design to esomeprazole (40 or 20 mg twice daily) with or without aspirin (300-325 mg/day) for \geq 8 years
- patients taking NSAIDs were excluded
- trial intentionally recruited fewer women because of lower risk of esophageal carcinoma in women with Barrett esophagus compared to men
- primary outcome was composite of all-cause mortality, esophageal adenocarcinoma, and high-grade dysplasia
- median treatment duration 8.9 years
- 99% completed study and were included in analysis
- comparing aspirin vs. no aspirin (both in combination with esomeprazole 40 or 20 mg twice daily)
 - high-grade dysplasia in 3.3% vs. 4.8% ($p = 0.053$, NNT 67)
 - primary outcome in 11.2% vs. 13.5% ($p = 0.068$, NNT 44)
 - men 11.7% vs. 14.4% ($p = 0.07$, NNT 32)
 - women 9.1% vs. 9.9% (not significant)
 - all-cause mortality 6.4% vs. 7.9% (not significant)
 - esophageal adenocarcinoma in 3.1% vs. 3.1% (not significant)
 - treatment-related serious adverse events in 1.3% vs. 0% (no p value reported)
- Reference - AspECT trial ([Lancet 2018 Aug 4;392\(10145\):400](#) [full-text](#))

This trial compared aspirin plus esomeprazole vs esomeprazole alone and found benefit with the combined treatment. Though the composite primary outcome was significantly improved with aspirin, DynaMed/DynaMedex chose to focus on its effect on the risk of high-grade dysplasia, the component of the composite outcome that drove the difference (there were no significant differences in the rates of the other components). The strength of the evidence was downgraded to level 2 (midlevel) evidence due to the lack of blinding of patients and clinicians.

Understanding and Appraising Studies About Diagnostic Tests

Unlike treatment questions, which seek to investigate a cause-and-effect relationship, diagnostic questions seek to study associations relevant for the diagnostic process. Therefore, when studying diagnostic tests, there is generally no need to look for a cause-and-effect relationship (and thus no need for an RCT), until the very late phase of diagnostic research, whereby one is interested in assessing the impact of adopting a diagnostic or screening procedure. Indeed, the implementation questions may require an RCT-driven answer. Before getting to that stage, however, all we are interested in is whether there is an association between a test (in the broadest possible meaning of lab or imaging test results, patient characteristics, or signs/symptoms) and a clinical condition.

The best way to study the utility of a test is in a cohort of patients who present a “diagnostic dilemma” (sometimes simply referred to as “diagnostic uncertainty,” the typical situation where the picture is not so clear-cut that you can “directly” be confident the disease is present or absent). This was briefly discussed in the section [Diagnosis Questions and Extending PICO](#). But why is this important? Why not study the test exclusively in people who have the condition we hope to identify or rule out? Because the role for a diagnostic test is to classify the subjects as positive (for example, diseased) and negative (for example, nondiseased). Consequently, to appraise how well a test performs in distinguishing positives and negatives, you have to apply it to a population comprised of both.

For instance, imagine a hypothetical scenario where a group of researchers is interested in investigating the possible diagnostic utility of PK543, a recently discovered blood protein, for diagnosing a pulmonary embolism. The researchers assemble a cohort of patients known to have pulmonary embolisms and test them all for PK543. Intriguingly, they report every patient in their study tested positive for PK543. Should we begin ordering this simple blood test to diagnose pulmonary embolism? Not so fast. What if PK543 is a blood protein present in everyone? This study fails to account for this. To see how ludicrous these results may end up being, replace all instances of “PK543” with “albumin.”

So, we must study the utility of tests in patients who present a “diagnostic dilemma.” In the example from the section [Diagnosis Questions and Extending PICO](#), this was the diagnostic utility of BNP in patients who present with undifferentiated respiratory symptoms. We would expect some, but not all, of these patients to have heart failure as the underlying cause of their symptoms. Therefore, if simplifying the BNP test into “positive” and “negative,” we can think of this cohort of patients posing a “diagnostic dilemma” as follows:

	Patients With Heart Failure (+HF)	Patients Without Heart Failure (-HF)
Test positive for BNP (Test +)	+HF Test+	-HF Test+
Test negative for BNP (Test -)	+HF Test-	-HF Test-

But wait, how do we know who has heart failure in the above table? This is why it is imperative to see the utility of any test be compared against the “gold standard” (sometimes simply referred to as the “reference standard”) for diagnosis. For heart failure, the gold standard for diagnosis is echocardiography. Therefore, in the hypothetical study leading to the table above, every patient must receive echocardiography to confirm or rule out the presence of heart failure.

Indeed, in addition to assessing whether the test under investigation was applied equally to all patients, key elements of appraising diagnostic studies are whether the gold standard was used and whether the gold standard was applied equally to all patients. Otherwise, the utility of the test in question cannot be soundly investigated. Of course, as we saw for studies on treatment questions, there are other aspects of appraising studies on diagnostic questions, but again, we will not attempt to provide an exhaustive consideration of such here.

Example Diagnostic Evidence Summary from DynaMed/DynaMedex

STUDY SUMMARY

focused cardiac ultrasound (FOCUS) may help rule out PE in adults with tachycardia or hypotension DynaMed Level 2

DIAGNOSTIC COHORT STUDY: Acad Emerg Med 2019 Nov;26(11):1211 [↗](#)

Details ^

- based on diagnostic cohort study with possible selection bias
- 136 adults (mean age 56 years) presenting to emergency department with suspected PE and tachycardia (heart rate \geq 100 beats/minute) or hypotension (systolic blood pressure $<$ 90 mmHg) had FOCUS and CT angiography (reference standard)
 - FOCUS included assessment of right ventricular dilation, McConnell sign (hypokinesis of right ventricle with apical sparing), septal flattening, tricuspid regurgitation, and tricuspid annular plane systolic excursion
 - FOCUS examination was considered positive if any component was abnormal and negative if all components were normal
- patients were enrolled as convenience sample when 2 FOCUS readers were available (not consecutive patients)
- 27.2% had PE by CT angiography
- diagnostic performance of FOCUS for detection of PE
 - overall
 - sensitivity 92%
 - specificity 64%
 - positive predictive value 49%
 - negative predictive value 95%
 - in subgroup of 98 patients with heart rate \geq 110 beats/minute, sensitivity was 100% and specificity was 63%
- FOCUS had high interobserver agreement (kappa = 1) based on 2 readers
- Reference - Acad Emerg Med 2019 Nov;26(11):1211 [↗](#)

In this diagnostic cohort study, a negative result on focused cardiac ultrasound appeared helpful for ruling out pulmonary embolisms in patients with tachycardia or hypotension, exhibiting high sensitivity and negative predictive value. The positive predictive value was relatively low, suggesting a positive result may be less informative. The strength of the evidence was downgraded to level 2 (midlevel) evidence due to the enrollment of a “convenience sample” of patients presenting when investigators were available. This is an

example of selection bias because an unknown number of eligible patients may have been excluded, potentially skewing the spectrum of included patients. Using a consecutive or random sample can help avoid this potential bias.

Understanding and Appraising Systematic Reviews and Meta-Analyses

Even among the most rigorous studies, there will always be some degree of inescapable variation in the data. This inescapable variation is termed random error. The amount of random error decreases as the cumulative amount of data pertaining to the question increases. This is known as the law of large numbers. However, even with a relatively larger study, it is often desirable to see the findings from that study replicated at least once due to how statistical inference works. Additionally, for many questions, there usually are multiple, smaller studies seeking to answer the same or a similar question.

Briefly, a systematic review is a research design that falls into the category of secondary literature (sometimes called secondary research), not because it is of lower quality or less desirable than primary literature, but rather, because it is a type of research that relies on primary literature being available. Systematic reviews seek to systematically identify and summarize all the available literature that is relevant to a targeted question and meets specified eligibility criteria for consideration. The systematic approach is key, and it helps reduce bias and increase reproducibility. In a word, it increases reliability. Therefore, assessing the methods used to identify the primary literature relevant to the question of interest is a key part of appraising a systematic review.

Eligibility criteria can vary among systematic reviews and assessing whether the eligibility criteria are appropriate is also a part of appraising systematic reviews. For instance, in seeking to answer a question about the effects of a treatment, some systematic reviews might only consider RCTs, whereas other systematic reviews might also allow cohort studies. Although including cohort studies might seem problematic based on the considerations outlined in the section [Understanding and Appraising Treatment Studies](#), we also saw in the same section that sometimes we may have to look to lower forms of evidence for insight (for example, assessing rare outcomes or addressing questions for which there is no RCT evidence).

It is common for a meta-analysis to be a part of a systematic review, but this is not always the case. Heterogeneity is often a key consideration in deciding whether to meta-analyze results. Heterogeneity speaks to the degree to which primary research studies differ from one another. They can differ in the type of populations studied or excluded, outcomes assessed, length of follow-up, interventions undertaken, and so on. There are a number of ways to assess and explore apparent heterogeneity, but in general, heterogeneity should

be anticipated to the fullest extent possible and explicitly handled when planning a systematic review.

For instance, consider a systematic review seeking to summarize the available evidence for the efficacy of NSAIDs in treating pain and disability related to osteoarthritis of the knee. We would ideally find evidence exclusively about patients with osteoarthritis of the knee. We might accept the inclusion of studies that had a certain percentage of patients with osteoarthritis of the hip or of another peripheral joint. However, we'd start to question the results if the authors also included studies that enrolled patients with rheumatoid arthritis and arthralgia from systemic lupus erythematosus. And we would be very confused by a decision to also include studies of patients with fibromyalgia, painful lower limb neuropathy secondary to diabetes, and reflex sympathetic dystrophy/complex regional pain syndrome.

In the first situation involving only patients with osteoarthritis of the knee, we would expect the clinical heterogeneity of the population to be relatively low. We would expect more heterogeneity if there were some patients with osteoarthritis of the hip or of another peripheral joint, but probably not to a problematic extent. However, once we cross into including patients with rheumatoid arthritis or any of the other conditions listed thereafter, we veer into a situation where the clinical heterogeneity of the population would be expected to be substantial. This can result in problems when trying to produce meaningful and reliable estimates for our original question (the efficacy of NSAIDs in treating pain and disability related to osteoarthritis of the knee).

Meta-analysis is the statistical aggregation of the data from the studies included in the systematic review. If the systematic review and meta-analysis are done well, the results obtained may offer a better quantitative estimate of the "truth" than the individual studies contributing data to the systematic review and meta-analysis (this again speaks to the law of large numbers). However, while it is generally true that a systematic review or meta-analysis is more reliable than the individual studies it includes, it would be overly simplistic to automatically consider systematic reviews and meta-analyses as high-quality evidence, as this assumes the underlying research was high-quality.

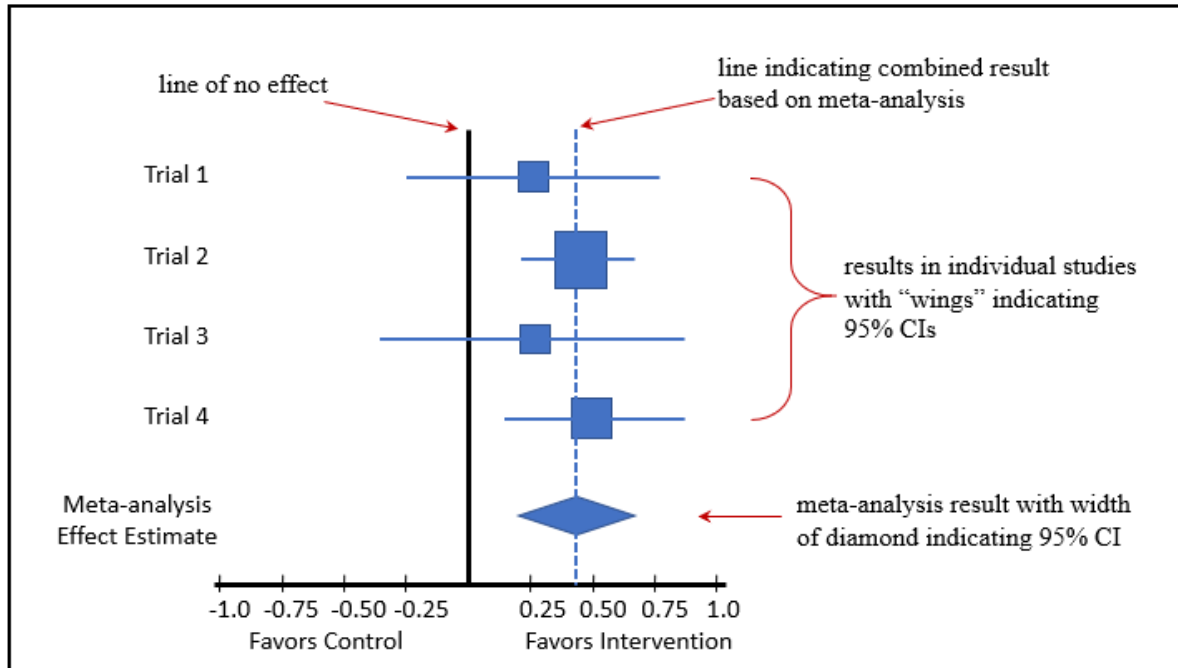
A well-done meta-analysis of poor-quality trials results in low quality and unreliable results from the meta-analysis. Similarly, a poorly conducted systematic review, such as one that combines clinically heterogeneous trials, with a meta-analysis of high-quality data does not automatically become a high-quality meta-analysis. Finally, sometimes a systematic review and meta-analysis may actually confuse things more than the individual RCTs on the matter. An example could be the question "Is enoxaparin more effective than unfractionated heparin in preventing recurrent cardiovascular events in patients with non-ST segment-elevation acute coronary syndrome (NSTEMI-ACS)?" Although there are systematic reviews and meta-analyses and even guideline statements on the matter, some

of these failed to account for how treatment of NSTEMI-ACS has evolved with time.¹¹ (Mayer, 2017) Therefore, meta-analyzing all the RCTs comparing enoxaparin to unfractionated heparin may not answer the question better than considering the individual RCTs (the evolution of care also represents an example of clinically important heterogeneity).

Let's take a closer look at forest plots. Forest plots offer a graphic representation of the individual study results and the weighted combined effect estimate, with weights for individual studies usually based on study size. The relative risks, odds ratios or hazard ratios of individual studies are represented by boxes, with the size of the box conveying the relative weight of that study result compared to the others being analyzed. The horizontal lines through the boxes represent the confidence interval for that study, with wider "wings" representing wider confidence intervals. A diamond at the bottom of the plot represents the combined result – the width of the diamond reflects the confidence interval of the combined estimate of effect.

Each of these results is plotted with respect to a solid vertical line at midline, called the line of no effect, which represents the null, or no difference between groups. The vertical dashed line projecting upward from the midpoint of the diamond is drawn to compare the combined study result to the individual results (boxes), which also helps to highlight heterogeneity among the included studies.

Heterogeneity among individual study results can be assessed statistically in a variety of ways. The simplest is the I^2 (I-squared) statistic. The higher the I-squared value, the higher the concern for significant heterogeneity that may affect the reliability of the combined effect estimate. There are no precise cut-off values for interpreting I-squared, but in general, an I-squared value of less than 25% is considered low heterogeneity and a value greater than 50% indicates substantial heterogeneity.



Example of a forest plot.

When performing a systematic review and meta-analysis, it is important for the internal validity of the analysis to avoid "publication bias." Publication bias can occur due to the tendency for editors and reviewers to not publish negative studies (those that don't find a difference between 2 interventions). Review authors should attempt to identify and include all available studies evaluating the research question, even those that did not get published, to ensure the study sample is not biased toward studies with positive effects. Publication bias can be visually assessed with a "funnel plot," which plots effect size on the horizontal axis vs effect precision on the vertical axis for each included study. Results from larger studies usually have greater precision and these are placed toward the top of the funnel plot, with results from smaller studies appearing lower down. When there is no publication bias, the points for individual studies in the plot will be roughly symmetrically distributed on either side of the effect estimate from the meta-analysis.

An asymmetric plot suggests the analysis includes a biased sampling of studies, and therefore, may not represent the most accurate effect estimate. An asymmetric funnel plot may be due to things other than publication bias. If the larger trials are clinically heterogeneous from the smaller ones, this may lead to asymmetry even if there is no publication bias. For funnel plots to be useful, there should typically be at least 10 trials in the meta-analysis.


Finally, there is discussion about whether so-called “mega-trials” (a nebulous term with respect to sample size, but conceptually signaling a very large RCT, often with more than 10,000 participants) might obviate systematic reviews and meta-analyses for the given question. For instance, consider the case of apixaban vs warfarin for the combined outcome of stroke and other systemic embolic events (hereafter, “outcome events”) in nonvalvular atrial fibrillation. The ARISTOTLE trial¹² was the pivotal RCT for this question, randomizing 18,201 patients and following them for a median of 1.8 years (with a total of 477 outcome events over the course of the study). However, in the relevant Cochrane review on the matter,¹³ ARISTOTLE was not the only RCT identified or included.¹⁴ The review authors also included the ARISTOTLE-J trial. However, ARISTOTLE-J was much smaller (222 patients randomized and a total of 3 outcome events), and thus only contributed 1.52% of the data for the meta-analytic estimate in the Cochrane review.


Perhaps more notably (and speaking to the importance of appraising the methods of a systematic review and meta-analysis), ARISTOTLE-J had the primary intent to study major and clinically relevant nonmajor bleeding. It was conducted exclusively in patients of Japanese heritage, it was very short (12 weeks), and it randomized 74 of the participants to a dose of apixaban that is only recommended in people who are both ≥ 80 years of age and ≤ 60 kg or who have a creatinine level ≥ 1.5 mg/dL and are either ≥ 80 years of age or ≤ 60 kg. Therefore, it is questionable whether the meta-analysis of the results of ARISTOTLE-J and ARISTOTLE serve us better in this situation compared to using only ARISTOTLE’s results, and although the results are quantitatively similar, the methodologic considerations of whether ARISTOTLE-J should be considered alongside ARISTOTLE apply, nonetheless.


Example Systematic Review Evidence Summaries from DynaMed/DynaMedex

STUDY SUMMARY

glucocorticosteroids may not reduce all-cause mortality up to 3 months in patients with alcoholic hepatitis DynaMed Level 2

COCHRANE REVIEW: [Cochrane Database Syst Rev 2019 Apr 9;\(4\):CD001511](#) 

Details 


- based on Cochrane review with wide confidence intervals
- systematic review of 16 randomized trials comparing glucocorticosteroids vs. placebo or no intervention in 1,861 patients aged 25-70 years with alcoholic hepatitis
 - glucocorticosteroids included prednisolone or 6-methylprednisolone given orally or parentally
 - median duration of treatment was 28 days
- 15 trials contributed data for meta-analysis
- no significant differences up to 3 months in
 - all-cause mortality (risk ratio 0.9, 95% CI 0.7-1.15) in analysis of 15 trials with 1,861 patients, but CI includes possibility of benefit or harm
 - liver-related mortality (risk ratio 0.89, 95% CI 0.69-1.14) in analysis of 15 trials with 1,861 patients
 - health-related quality of life in 1 trial with 377 patients
 - risk of serious adverse events in analysis of 15 trials with 1,861 patients
- Reference - [Cochrane Database Syst Rev 2019 Apr 9;\(4\):CD001511](#) 

This systematic review and meta-analysis of glucocorticoids for alcoholic hepatitis was assessed as providing level 2 (midlevel) evidence due to imprecision in the effect estimate for mortality. Though there was no significant difference in the mortality between glucocorticoid and placebo groups, the confidence interval for the risk ratio was too wide to exclude the potential benefit or harm with glucocorticoids.

STUDY SUMMARY

bariatric surgery may reduce urinary albumin excretion in adults with diabetic kidney disease DynaMed Level 3SYSTEMATIC REVIEW: [Surg Obes Relat Dis 2016 Jun;12\(5\):1037](#) 

Details ^

- based on systematic review of observational studies without clinical outcomes
- systematic review of 15 observational studies (12 cohorts, 2 case-controls, and 1 cross-sectional study) evaluating urinary albumin excretion before and after bariatric surgery in 993 adults aged ≥ 18 years with diabetic kidney disease
- diabetic kidney disease defined as increased urinary albumin excretion in absence of other renal disease
 - microalbuminuria if urinary albumin:creatinine ratio 30-300 mg/g of creatinine
 - macroalbuminuria if urinary albumin:creatinine ratio > 300 mg/g of creatinine
- bariatric procedures were Roux-en-Y gastric bypass, gastric banding, vertical banded gastroplasty, or biliopancreatic diversion
- follow-up ranged from 12 to 120 months
- bariatric surgery associated with reduction in
 - urinary albumin:creatinine ratio (mean difference -6.6 mg/g, 95% CI -9.19 to -4.02 mg/g) in analysis of 8 studies, results limited by significant heterogeneity
 - albuminuria (mean difference -55.76 mg/24 hours, 95% CI -92.11 to -19.41 mg/24 hours) in analysis of 6 studies, results limited by significant heterogeneity
- Reference - [Surg Obes Relat Dis 2016 Jun;12\(5\):1037](#) 

This systematic review includes data from observational studies only because no RCTs addressing the clinical question at hand were identified in the systematic search. It was assessed as providing level 3 (lacking direct) evidence, because the outcome measure is a surrogate or nonclinical outcome that does not necessarily correlate with improvement in patient-oriented (clinical) outcomes.

X. Guidelines and Recommendations

Clinical practice guidelines can provide a source of professional support for clinical decision-making, can lessen the amount of work we have to do to arrive at a decision, and can even reduce medical-legal worries for some. However, guidelines can vary tremendously in quality, come from many different sources, and some have little to no

evidence to back them up. In some cases, guidelines can acknowledge the best available evidence and yet make opposing recommendations. So, just as it's important to be able to understand what factors make treatment and diagnostic studies more or less trustworthy, it's equally, or perhaps more important, to be able to do the same with guidelines.

Where Do Clinical Practice Guidelines Come From?

A clinical practice guideline aims to guide decisions by providing criteria regarding diagnosis, management, and treatment in specific areas of health care. These come from:

- Government agencies (for example, The US Preventive Services Task Force)
- Medical associations (for example, The American Thoracic Society)
- Payers/insurers (for example, The National Institute for Health and Care Excellence [NICE])
- Other organizations (for example, The Global Initiative for Chronic Obstructive Lung Disease)
- Foundations (for example, The National Osteoporosis Foundation)
- Advocacy groups (for example, The American Diabetes Association)
- Commercial organizations
- Ad hoc groups, often funded by the pharmaceutical industry
- Local health care systems

There are 3 types of guidelines: consensus-based, evidence-based, and evidence-linked.

Consensus-Based Guidelines

These are produced by expert panels. They're developed by bringing together a group of experts who decide how to write the guideline, often under the auspices of a professional society. Evidence is likely used by the group in some way, but there is no indication in the guideline regarding how they found, evaluated, and interpreted the evidence.¹⁵ As a result, the quality of the recommendations is low.¹⁶

Evidence-Based Guidelines

These are often based on a systematic review of the literature, but they lack an explicit, transparent, and reproducible validity assessment of the evidence considered when making recommendations. Sometimes explained as a panel of experts who state, "trust us, we've reviewed the literature," evidence-based guidelines ask clinicians to be assured the writers have used the evidence appropriately.

Evidence-Linked Guidelines

These reflect the highest level of validity in guidelines and can be identified by a methods section explaining how the evidence was selected and summarized, how it was graded, and how it was used to inform the decision-making behind the guidelines. Frequently, this type of guideline has 2 documents: 1 document contains the recommendations that are linked to a second document that reports the evidence supporting these guidelines.

Potential Problems with Guidelines

Guidelines can differ markedly in their quality. An independent review showed that less than half of the 130 guidelines evaluated by the authors met more than 50% of the requirements for good guideline development.¹⁸ Problems with guidelines include:

- **Lack of transparency.** There should be a clear and reproducible path from the recommendations back to the evidence.
- **Guidelines that don't guide.** Sometimes, recommendations are written in a general form that makes it hard for clinicians to understand exactly what is being recommended. For example, a depression guideline suggested that "it is not unreasonable to try exercise in certain patients."
- **Lack of systematic review of the literature.** As a result, guidelines sometimes suffer from selective citing of research that supports a certain position, ignoring research that doesn't support this position.
- **Conflicts of interest.** These can be financial, intellectual, or professional.
 - A **financial** conflict of interest can occur when a guideline developer receives research funding by a manufacturer of a product affected by the guideline, or if they are a paid speaker or consultant to a manufacturer or entity. These financial arrangements may result in a conscious or unconscious bias.
 - **Intellectual** conflicts occur when one's research or experiences unduly influence one's ability to evaluate other types of evidence. It produces a sort of "tunnel vision."
 - A **professional** conflict of interest occurs when a guideline is sanctioned or approved by a professional society. Since these groups exist to support and advocate for their members, there is a risk the needs of the groups' members (for example, the specialty group) supersedes the needs of the patients they serve, since a professional society has a primary responsibility to promote its members' interests.

Critical Appraisal of Guidelines and Recommendations

Guidelines from different organizations often include recommendations that can substantially differ from one another. For example, as of 2020, the American Diabetes Association recommends a hemoglobin A1C goal of under 7% for nonpregnant adults, while the American Association of Clinical Endocrinologists recommends a target of under 6.5%, and the American College of Physicians recommends 7%-8%. As a result, it is important to be able to evaluate a guideline by assessing its relevance and utility, trustworthiness, and any influences on the interpretation of the evidence to develop recommendations. Tools like the G-TRUST checklist¹⁷ provide a framework for such assessment. The G-TRUST checklist incorporates the following factors for assessing relevance and utility, trustworthiness, and influences on interpretation:

Relevance and utility can be evaluated by determining whether:

- The recommendations focus on improving patient-oriented outcomes, not disease-oriented outcomes, explicitly comparing demonstrated benefits vs harms to support clinical decision-making.
- The recommendations are clear and actionable.
- The patient populations and conditions are relevant to your clinical setting.

Trustworthiness can be determined by assessing whether:

- The guidelines are based on a systematic review of the research data, usually through linking the recommendations to a systematic review of the available literature.
- The recommendations important to you are based on graded evidence and include a description of the quality of the evidence (for example, strong, weak).
- The guideline development group includes someone with methodologic expertise, such as a statistician or epidemiologist.

Influences on interpretation can be assessed by determining whether:

- The chair of the guideline development committee and a majority of the rest of the committee are free of declared financial conflicts of interest, and the guideline development group did not receive industry funding for developing the guideline.
- The guideline development group includes members from the most relevant specialties (including primary care physicians) and includes other key stakeholders, such as patients, payer organizations, and public health entities, when applicable.

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) System

The GRADE system is the result of an international collaboration to develop a transparent approach to grading quality (or certainty) of evidence and strength of recommendations. Many guideline development groups use this system. DynaMed/DynaMedex uses the GRADE system to label synthesized recommendations as strong or conditional based on the strength of the evidence and a balancing of the positive and negative aspects of the intervention on patients' lives.

- Strong recommendations are used when, based on the available evidence, clinicians (without conflicts of interest) consistently have a high degree of confidence that the desirable consequences (health benefits, decreased costs and burdens) do, or do not, outweigh the undesirable consequences (harms, costs, burdens).
- Weak recommendations are used when, based on the available evidence, clinicians believe desirable and undesirable consequences are finely balanced, or appreciable uncertainty exists about the magnitude of expected consequences (benefits and harms). Weak recommendations are used when clinicians disagree in judgments of relative benefit and harm or have limited confidence in their judgments. Weak recommendations are also used when the range of patient values and preferences suggests informed patients are likely to make different choices.

More information about the GRADE system can be found at gradeworkinggroup.org.

References

1. Lind J. A treatise of the scurvy. 1753.
<https://archive.org/details/b30507054/page/n6/mode/2up>
2. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet*. 2017;390(10092):415-423. doi:10.1016/S0140-6736(16)31592-6
3. Guyatt GH. Evidence-based medicine. *ACP J Club*. 1991;114:A16. doi:10.7326/ACPJC-1991-114-2-A16
4. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ*. 1996;312(7023):71-72. doi:10.1136/bmj.312.7023.71
5. Ebell MH, Shaughnessy AF, Slawson DC. Why are we so slow to adopt some evidence-based practices? *Am Fam Physician*. 2018;98(12):709-710.
6. Densen P. Challenges and opportunities facing medical education. *Trans Am Clin Climatol Assoc*. 2011;122:48-58.
7. Alper BS, Hand JA, Elliott SG, et al. How much effort is needed to keep up with the literature relevant for primary care? *J Med Libr Assoc*. 2004;92(4):429-437.
8. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *Evid Based Med*. 2016;21(4):125-127. doi:10.1136/ebmed-2016-110401
9. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med*. 2007;356(24):2457-2471. doi:10.1056/NEJMoa072761
10. Teo KK, Yusuf S, Furberg CD. Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction. An overview of results from randomized controlled trials. *JAMA*. 1993;270(13):1589-1595.
11. Mayer M. Anticoagulants in ischemia-guided management of non-ST-elevation acute coronary syndromes. *Am J Emerg Med*. 2017;35(3):502-507. doi:10.1016/j.ajem.2016.12.070
12. Granger CB, Alexander JH, McMurray JJ, et al; ARISTOTLE Committees and Investigators. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med*. 2011;365(11):981-992. doi:10.1056/NEJMoa1107039
13. Ogawa S, Shinohara Y, Kanmuri K. Safety and efficacy of the oral direct factor xa inhibitor apixaban in Japanese patients with non-valvular atrial fibrillation. -The ARISTOTLE-J study-. *Circ J*. 2011;75(8):1852-1859. doi:10.1253/circj.cj-10-1183
14. Bruins Slot K Mh, Berge E. Factor Xa inhibitors versus vitamin K antagonists for preventing cerebral or systemic embolism in patients with atrial fibrillation. *Cochrane Database Syst Rev*. 2018;3(3):CD008980. doi:10.1002/14651858.CD008980.pub3

15. Ioannidis J PA. Professional societies should abstain from authorship of guidelines and disease definition statements. *Circ Cardiovasc Qual Outcomes*. 2018;11(10):e004889. doi:10.1161/CIRCOUTCOMES.118.004889
16. Sinuff T, Patel RV, Adhikari NKJ, Meade MO, Schünemann HJ, Cook DJ. Quality of professional society guidelines and consensus conference statements in critical care. *Crit Care Med*. 2008;36(4):1049-1058. doi:10.1097/CCM.0b013e31816a01ec
17. Kung J, Miller RR, Mackowiak PA. Failure of clinical practice guidelines to meet institute of medicine standards: two more decades of little, if any, progress. *Arch Intern Med*. 2012;172(21):1628-1633. doi:10.1001/2013.jamainternmed.56
18. Shaughnessy AF, Vaswani A, Andrews BK, et al. Developing a clinician friendly tool to identify useful clinical practice guidelines: G-TRUST. *Ann Fam Med*. 2017;15(5):413-418. doi:10.1370/afm.2119

EBM Glossary

Absolute Risk

The probability of developing a condition or having a specific event during a given period of time. In a clinical study, absolute risk is the ratio of the number of study participants in a group having an outcome event of interest during a specified follow-up time to the total number of participants in the group. Risk can refer to both adverse and favorable outcomes (for example, risk of incident heart failure or risk of symptom resolution).

Absolute Risk Difference

The arithmetic difference in the risk of an outcome between 2 groups (for example, intervention and control groups are usually expressed as intervention - control). May be referred to as absolute risk reduction (ARR) if the intervention is beneficial (higher rate of favorable or lower rate of adverse outcomes) or as absolute risk increase (ARI) if the intervention is harmful.

Allocation Concealment

Blinding of a trial's randomization sequence to the study personnel responsible for enrolling participants, so the specific group to which a patient will be randomized is unknown at the time of enrollment. This process is separate from the blinding of patients and investigators during the conduct of the trial. Lack of allocation concealment can lead to biased assignment to trial groups, which can affect the reliability of the study's findings. For instance, a researcher who believes an unproven intervention is beneficial might be inclined to avoid enrolling the sickest patients to the intervention group for the fear they would do poorly anyway. This could result in the control group having sicker patients, and the benefit seen with the treatment would be falsely amplified by having more patients likely to do well regardless of the intervention.

Attention Control

A form of placebo control in a randomized trial intended to balance the time and attention received from study personnel by participants in the intervention group, when a drug placebo is inappropriate. For example, an educational session on general health concerns could act as an attention control for a brief counseling intervention.

Attrition Bias

Type of selection bias whereby a meaningful percentage of participants withdraw or are otherwise lost from the study (generally more than 20% overall or in at least one group), introducing potential imbalances in the characteristics of the groups.

Background Question

General questions regarding a topic (for example, “what is the disorder?” “What causes it?” “How does it present?” “What are some management options?”). These questions can be answered by using “background” resources, such as text references and narrative reviews in journals, which give a general overview of the topic.

Blinding

Experimental method used to ensure the people involved in a study do not know what intervention is being given to any specific individuals or groups during the trial. Blinding “people involved in the study” could mean a single-blind mechanism (either the researchers or the patients are blinded, but not both) or a double-blind mechanism (both the researchers and the patients are blinded). Blinding, also called “masking” is critical for distinguishing the true benefit of the intervention from the placebo effect. Blinding helps ensure the groups in the trial have experiences that are as identical as possible aside from the intervention(s) of interest, including managing concomitant treatments and diagnostic procedures, controlling for expectations and the meaningfulness of receiving or not receiving a given intervention. If neither patients nor researchers are blinded, a trial is referred to as “open label.” Some open-label trials include a blinded independent group of outcome assessors to reduce the potential bias resulting from knowledge of group assignment.

Case-Control Study

A retrospective study in which patients with an outcome of interest, such as a specific disease diagnosis (cases), and individuals without the diagnosis (controls) are compared for previous exposures to assess the association between exposure and disease. This design is often used to identify risk factors for certain diseases, such as cancer, or to study possible rare side effects of certain treatments.

Case Finding

An approach to identifying disease before symptoms emerge directed at at-risk individuals, as opposed to screening, which is directed at the general population. Case finding emphasizes diagnosing the disease while screening is concerned with both detecting and ruling out disease.

Case Series

An observational interventional study reporting outcomes for patients who have all received the same treatment. Case series studies may be prospective or retrospective, and always lack a control group. DynaMed/DynaMedex conclusions based on uncontrolled trials always receive a level of evidence 3 due to the lack of a comparison group.

Citation Bias

Type of bias that occurs when negative studies are less likely than positive studies to be cited as previous or established evidence. This can give a false impression of the benefits of an intervention. Citation bias is related to publication bias wherein negative studies are less likely to be published.

Clinical Practice Guideline

Document prepared by a guideline committee that aims to guide decisions by providing explicit statements regarding how to diagnose and treat specific clinical conditions.

Clinically Importance Difference

The concept of whether a difference between 2 groups demonstrated in study result matters in a clinical context. This is different from statistical significance. A finding can be statistically significant while having little, no, or uncertain clinical importance. For example, an absolute risk reduction for myocardial infarction by 0.2% over 5 years with a new drug would not be considered clinically important even if it's statistically significant. A "minimal clinically important difference (MCID) for an outcome measure is the smallest difference in that measure that leads to a noticeable effect for the patient.

Clinical Outcome

Study outcome measures (effects of an intervention) that have a direct impact on a patient's well-being, including symptoms, functional ability, or quality of life, or any outcome, such as a disease diagnosis, that is directly attributable to symptoms. Also referred to as "patient-oriented" outcomes.

Cohort Study

An observational study in which a group of participants with an exposure to a specific risk factor or intervention are compared to a group without that exposure and followed for outcomes to assess the association between exposure and outcome. Cohort studies can be either prospective, in which participants are enrolled prior to an outcome event, or retrospective, in which both exposures and subsequent outcomes have been recorded prior to study initiation.

Confidence Interval

A statistic providing a range of results that could be expected if a study were repeated. This range is typically set at 95%. The interval includes the mean values considered statistically compatible with the data in the study (as long as the assumptions of the model used to calculate the confidence interval are also met) and gives insight into the precision of the results. The width of the confidence interval is a function of the sample size and variability within the sample.

Confounding

Type of bias that occurs due to the inappropriate attribution of an effect to an intervention, when in fact there is a third contributing factor that is (at least in part) responsible for the observed effect.

Continuous Outcome

An outcome measure with values that lie on a continuum. Examples include body weight, BP, duration of hospitalization, or pain measured on a 0- to 100-mm visual analog scale. In study results, continuous outcomes are generally expressed by a measure of central tendency, such as the mean or median, for each study group. Efficacy of an intervention is usually expressed as the arithmetic difference in the average values between groups.

Control Group

A group in a study that does not receive the intervention of interest. This can include no treatment, placebo, standard of care, alternative treatment, wait list control, or other control.

Correlation

A statistical measure that quantifies the strength of association between 2 factors (for example, a risk factor and incidence of a disease). Correlation values can be positive or negative within the range -1 to 1. A positive correlation indicates an increasing level of the risk factor is associated with increased risk of disease, and a negative correlation indicates the opposite relationship. While strong correlation can be predictive of increased risk, it does not imply a causal relationship between the risk factor and the disease.

Credible Interval

A Bayesian statistics measure of the precision of an effect estimate analogous to the confidence interval, providing a range of values with a specified probability (usually 95%) of containing the true population effect.

Critical Appraisal

The process of carefully and systematically assessing research publications for their degree of trustworthiness. More information about DynaMed/DynaMedex's critical appraisal process can be found at www.elsevier.com/clinical-decisions/dynamed-solutions/about/evidence-based-process/methodology.

Detection Bias

Type of bias that can occur when there are systematic differences between groups in how outcome are determined. This can be due actual differences in assessment tools used in each group or due to investigator expectations. The latter bias can be minimized by

blinding of the outcome assessors with respect to the treatment the study participant received.

Dichotomous Outcome

A categorical outcome measure with only 2 possible states: patients either have an outcome event or they don't. In study results, dichotomous outcomes are expressed as the proportion of patients in a study group that have the event and provide an estimate of absolute risk. Differences in dichotomous outcomes between study groups may be described either in terms of absolute risk differences or relative risks.

Diagnostic Study

A study that reports the performance of a diagnostic test for detecting a condition of interest. A diagnostic cohort study assesses the agreement of positive and negative test results with the results of a trusted reference standard for the condition to estimate the sensitivity, specificity, and positive and negative predictive values of the test in a group of patients with a suspected disease. A diagnostic case-control study assesses the test's sensitivity for the disease in patients known to have the diagnosis, and the test's specificity in patients known not to have the diagnosis. See [Sensitivity](#), [Specificity](#), and [Predictive Values](#).

Evidence-Based Medicine

The judicious and intentional use of the best available evidence to inform health care decisions with the recognition there is a hierarchy of the quality of evidence and some types of evidence are more likely to represent the truth than other evidence.

Evidence-Informed Decision-Making

Finding, analyzing, and using evidence integrated with patient values and preferences to inform health care decisions.

False Negative

An incorrect negative result on a diagnostic test for a patient who does have a disease based on a reliable reference standard for diagnosis.

False Positive

An incorrect positive result on a diagnostic test for a patient who does not have a disease based on a reliable reference standard for diagnosis.

Foreground Question

A question about management (diagnosis, prognosis, treatment, etc.) of a specific patient, suitable to be answered with experimental evidence.

GRADE System

The Grading of Recommendations Assessment, Development, and Evaluation System is the result of an international collaboration to develop a transparent approach to grading quality (or certainty) of evidence and strength of recommendations. The GRADE system is used for recommendation statements in DynaMed/DynaMedex, but evidence grading for individual study summaries in DynaMed/DynaMedex uses a simple 3-tiered rating system based on our critical appraisal process. See [SORT](#).

Gold Standard

The generally agreed-upon most accurate mechanism to establish if a disease is present. Sometimes, the gold standard may be impossible to obtain while a patient is alive (for example, autopsy) or may be invasive (for example, pulmonary angiography). See [Reference Standard](#).

Hazard Ratio

The risk of an outcome (benefit or harm) with one treatment compared with another in a time-to-event analysis or survival analysis. Like the risk ratio, the hazard ratio is a measure of relative risk, but the risk ratio is based on cumulative data over a set period of time (for example, at follow-up of 1 year), while the hazard ratio represents the relative risk differential between 2 hazards, such as the function describing the risk at each time point over a period.

Heterogeneity

Heterogeneity refers to differences among the studies included in a systematic review or meta-analysis, which may limit the validity of pooled results. “Clinical heterogeneity” includes differences in patient populations, study methods, and outcomes assessed in the included studies, and its appraisal in DynaMed/DynaMedex is based on editorial judgment. “Statistical heterogeneity” is quantitatively assessed by comparison of the individual study results included in a meta-analysis.

Intention-to-Treat Analysis (ITT)

Analysis that includes all participants randomized to the study groups. It has specific methods for addressing noncompliance, withdrawal, and protocol deviations that include all of those randomized in the outcomes analysis. It maintains the prognostic balance offered by the original randomization. For positive outcomes (for example, cure or symptom resolution), an ITT analysis tends to generate conservative effect estimates. See [Per-Protocol Analysis](#).

Lead-Time Bias

A potential bias in screening studies arising from earlier detection of asymptomatic disease. Early detection may inflate a patient’s apparent survival time after diagnosis

compared to a patient diagnosed at the onset of symptoms even if screening has no effect on the time of death. Lead-time bias can overestimate of efficacy of screening for increasing survival.

Length Bias

Also called length-time bias. A potential bias in screening studies whereby the likelihood of detecting asymptomatic disease through screening tests is higher for indolent disease than for aggressive disease. (Aggressive disease has a higher risk of becoming symptomatic between screening tests.) Since the prognosis is likely to differ between indolent and aggressive disease, the apparent effect of screening on long-term outcomes is exaggerated. Overdiagnosis is an extreme form of length bias, in which screening detects disease that is so indolent it would not have been clinically apparent during a patient's lifetime.

Level of Evidence

A hierarchy applied to evidence that takes into account relevance, study design, and validity. Evidence in individual study summaries in DynaMed/DynaMedex is graded using a simple 3-tiered rating system based on our critical appraisal process. See [SORT](#). More information about DynaMed/DynaMedex's level-of-evidence hierarchy can be found in this [Levels of Evidence PDF](#).

Likelihood Ratio

A measure expressing the extent to which a test result or clinical finding increases or decreases the probability of a disease being present.

- Positive Likelihood Ratio (PLR or LR+): the probability of a positive test result in patients with the disease divided by the probability of a positive result in patients without the disease (Sensitivity / [1-Specificity])
- Negative Likelihood Ratio (NLR or LR-): the probability of a negative test result in patients with the disease divided by the probability of a negative result in patients without the disease ($[1 - \text{Sensitivity}] / \text{Specificity}$)

As a rule of thumb, a positive result on a test is considered highly predictive for ruling in disease if the test's PLR is greater than or equal to 10 and a negative result highly predictive for ruling out disease if the test's NLR is less than or equal to 0.1. A test's diagnostic utility decreases as the PLR and NLR approach a value of 1.

Loss to Follow-Up

A potential source of bias due to missing data from participants originally enrolled into a study who are not available at the time of data collection or have withdrawn from the study. Loss to follow-up can disrupt the balance between study groups and can limit the

validity of the results if the proportion lost is large relative to absolute treatment effect, if the loss rates differ between groups, or if the overall loss is more than 20%.

Masking

See [Blinding](#).

Meta-Analysis

A statistical technique for combining the findings from multiple independent studies, each addressing the same direct comparison, to synthesize single pooled effect estimates. This analysis is usually performed in conjunction with a systematic review and provides a more rigorous answer than simply collating study results (“4 studies say it works, 2 studies say it doesn’t, so I guess it works.”)

Network Meta-Analysis (NMA)

A set of statistical techniques incorporating both direct and indirect comparisons from multiple studies to assess comparative efficacy and safety among multiple treatment options. NMAs can provide pairwise comparisons between treatments that have not been directly compared in individual randomized control trials, but their primary purpose is to use the greatest amount of data possible to provide more robust effect estimates than may be possible using direct comparisons alone, as in standard meta-analyses.

Nonclinical Outcome

Study outcomes measuring disease state markers, such as BP or lab values, that may be important indicators of health status, but do not directly affect a patient’s quality of life. Also referred to as “surrogate” or “disease-oriented” outcomes. See [Clinical Outcome](#).

Number Needed to Treat (NNT)

The NNT tells us how many people need to be treated instead of not treated, or the number that need to be treated with 1 therapy instead of another, for 1 additional person to benefit over a certain period of time. The NNT is calculated as the reciprocal of the Absolute Risk Difference between groups.

Number Needed to Harm (NNH)

The NNH tells us how many people need to be treated instead of not treated, or the number that need to be treated with 1 therapy instead of another, for 1 additional adverse effect to occur over a certain period of time. The calculation is the same as for the number needed to treat.

Observational Study

A general term used to refer to most types of primary research that are not randomized, controlled trials. In DynaMed/DynaMedex evidence summaries, interventional conclusions

based on observational studies generally cannot receive a level of evidence 1 due to the potential bias introduced by the lack of randomization.

Odds

The ratio of the number of study participants in a group having an exposure or outcome event to the number of participants in the same group who don't have the exposure or event.

Odds Ratio

The ratio of the odds of an event between 2 study groups. A value of 1 indicates no difference between groups, values greater than 1 indicate increased odds of having an event in 1 group, and values less than 1 indicate reduced odds. The odds ratio is frequently misinterpreted as a measure of relative risk (or risk ratio). When the event rates are low (less than 10%), the odds ratio closely approximates the risk ratio, but as event rates increase, the odds ratio tends to overestimate the relative risk (the odds ratio is further from 1 in either direction than the risk ratio). See [Risk Ratio](#).

Overdiagnosis

Detection of a disease at an early asymptomatic stage through screening that would not have become symptomatic during a patient's lifetime and would not have been diagnosed clinically. Overdiagnosis can lead to unnecessary or even harmful treatment, such as surgery or radiation. Overdiagnosis cannot be identified in an individual patient since their future course is unknowable. It is detected by examining population data. See [Length Bias](#) and [Lead-Time Bias](#).

Performance Bias

Type of bias that occurs when participants or outcome assessors change behavior based on knowledge or perceived knowledge of treatment resulting in systematic differences between groups.

Per-Protocol Analysis

Comparison of treatment groups that limits analysis to only those who completed the treatment to which they were originally allocated. A per-protocol analysis ignores data from participants who withdrew or had major protocol deviations and typically results in a larger effect than intention-to-treat analysis. In superiority trials, this can be a source of bias, indicating a significant effect when none exists. However, for noninferiority trials, per-protocol analysis is the more conservative approach and is actually encouraged (usually in addition to intention-to-treat analysis). See [Intention-To-Treat Analysis \(ITT\)](#).

PICO

An acronym for **P**opulation (patient)/**I**ntervention (exposure)/**C**omparison (control)/**O**utcome. PICO provides a framework for focusing practical and relevant clinical questions.

Predictive Values

Predictive values are measures of diagnostic test performance estimating the probability for an individual patient of having or not having a target disease after either a positive or negative result on a diagnostic test.

- Positive Predictive Value is the probability a positive test result correctly identifies patients with the target disease.
- Negative Predictive Value is the probability a negative test result correctly identifies patients without the target disease.

In conjunction with disease prevalence, predictive values are potentially more informative at the point of care than test sensitivity and specificity. Prevalence is a measure of “pretest probability,” the probability of a patient having the disease prior to any diagnostic work-up. The predictive values provide “posttest probabilities” for that patient of having the disease based upon the results and are sometimes referred to as clinical performance parameters of a test. Sensitivity and specificity are sometimes referred to as “technical” performance parameters: they indicate the overall accuracy of a test in a population, but they provide little help for interpreting test results for individual patients. The predictive values of a test are dependent upon disease prevalence (positive predictive value goes up and negative predictive value goes down as prevalence increases), so they cannot be calculated from case-control study data, from which prevalence cannot be estimated.

Primary Research (Primary Literature)

An original study on a question that generates new data. Randomized controlled trials, cohort studies, and case-control studies are all examples of primary research.

Prognosis

An estimate of the likely course and outcome of a disease.

Publication Bias

The likelihood that negative results, such as studies that do not show a difference or benefit of a treatment, are less likely to be published.

P Value

A p value is a number between 0 and 1 that is used in statistics to gauge the statistical significance of a research finding. It is inversely proportional to the size and/or consistency of an effect of the research finding. The p value cutoff typically used to signal statistical

significance is less than 0.05, but this convention is arbitrary. In the conventional hypothesis testing statistical framework, the null hypothesis states there is no “true” difference in the outcome measure between experimental groups (for example, intervention and control), and that any observed difference is due to chance (sampling error). The p value indicates the probability you would obtain the observed difference (or an even bigger difference) if the null hypothesis were true. When the p value is below the cutoff, the null hypothesis is rejected in favor of the alternative hypothesis, that the intervention is responsible for the difference in the outcome. Reporting the confidence interval (CI) along with the p value is often preferred because the CI offers additional information not conveyed by the p value in isolation. See [Confidence Interval](#).

Post-Hoc Analysis

Statistical analysis that is not prespecified, therefore determined after data is already known and at high risk of bias.

Quasi-Randomized Trial

An interventional study in which participants are assigned to treatment groups by a nonrandom method (for example, by birth date, day of the week, or medical record number). Study summary conclusions in DynaMed/DynaMedex from quasi-randomized trials cannot receive a level of evidence 1 because the lack of true randomization may cause imbalances between groups.

Randomized Trial (also Randomized Controlled Trial or RCT)

- A study in which study subjects have an equal likelihood of being assigned to each study arm. RCTs offer a prognostic equivalence between groups that allows for identification of causal relationships to the independent variable being studied. This study design has less systematic bias than observational study designs. Most randomized trials are “parallel,” with participants concurrently treated in all study arms, and designed to assess the superiority of an intervention vs a control or alternative intervention. Crossover randomized trial: A trial in which participants are randomized to the order in which they receive all study treatments, rather than receiving just a single treatment. Participants serve as their own controls in paired statistical analysis.
- Cluster randomized trial: A trial in which groups of participants are randomized together based on membership in particular clusters (for example, patients in hospital wards, residents of a village). Each participant within a given cluster receives the same randomized treatment.
- “Stepped wedge trial” :A cluster randomized trial in which each cluster is randomized to the time of starting a treatment – each cluster has both control and treatment periods.

- **Noninferiority randomized trial:** A randomized trial designed to test whether the efficacy of a new intervention is within a predefined acceptable margin of the efficacy of an existing treatment, even if it may be slightly less effective overall, based on one limit of the confidence interval around the effect measure (lower limit for positive outcomes, upper limit for negative outcomes). The null hypothesis is that the new intervention is inferior. A statistically significant result (“noninferiority met”) suggests the efficacy of the new intervention is within the defined margin at the limit of the confidence interval. An “equivalence trial” is similar to a noninferiority trial, but tests whether the difference in efficacy between 2 interventions lies within predefined margins at both the upper and lower limits of the confidence interval.

Reference Standard

The method used in a diagnostic cohort study to establish whether a disease is present or absent, against which the performance parameters (sensitivity, specificity, predictive values) of an investigational diagnostic test are assessed. The reference standard used in a diagnostic study may not always be the gold standard for diagnosis, due to practical considerations. For example, a study might use CT as the reference standard for coronary artery stenosis, though invasive angiography would be expected for definitive diagnosis. See [Gold Standard](#).

Regression Analysis

A statistical modeling technique used to quantify the relationship between an outcome measure (for example, death) and one or more independent explanatory factors (for example, treatment received, age, comorbidities). A univariate regression uses only a single explanatory factor and may overestimate the effect of that factor on the outcome. A multivariate regression can account for possible confounding due to other factors. For example, a drug may appear to have a strong protective effect against death if analyzed in a univariate regression, but that effect may be diminished when the effects of other variables, such as the underlying condition the medication was prescribed for, are taken into account. The drug effect is called the crude estimate in the univariate analysis and the adjusted estimate in the multivariate analysis.

Risk Ratio (also Relative Risk or RR)

The ratio between outcome rates in the intervention and control groups, representing the relative benefit or harm associated with the intervention. Values greater than 1 indicate increased risk, and values less than 1 reduced risk. When the confidence interval around the relative risk estimate includes 1, then there is no statistical difference between the 2 treatments. Risk ratios (and other ratio-based measures, like odds ratios) are frequently reported in research studies, particularly in systematic reviews. Interpreting the clinical

importance of a significant difference in risk ratio can be complicated if presented without information about absolute risk difference. See [Absolute Risk](#) and [Absolute Risk Difference](#).

SORT (Strength of Recommendation Taxonomy)

A system for grading evidence based on individual studies and recommendations based on bodies of evidence designed to be easy to use at the point of care. It uses a simple 3-tiered rating hierarchy to evaluate the quality, risk of bias, consistency, and volume of evidence, and emphasizes the importance of patient-oriented outcomes for clinical decision-making. The SORT system provides the basis of DynaMed/DynaMedex's levels of evidence and critical appraisal processes.

Screening Test

A test that aims to identify disease in an asymptomatic individual. The most appropriate conditions for which to screen have a critical point before symptoms have begun at which point initiation of treatment would prolong health (morbidity or mortality) compared to initiating treatment after symptoms are present.

Secondary Research (Secondary Literature)

Research that appraises and/or summarizes primary research (primary literature). Systematic reviews and meta-analyses are perhaps the best examples of secondary research. Of note, "secondary" does not signify lower quality or desirability. Rather, it speaks to the fact that secondary research relies on the availability of primary research.

Secondary Analysis

A publication that reexamines data from a previously published study to report on additional outcomes from the original follow-up period that were not previously addressed or updated outcomes based on longer follow-up.

Selection Bias

Category of bias that can occur when selection of individuals, groups, or data for analysis is not equal or representative between groups. There are many more specific types of bias that fall under the category of selection bias, including sampling bias, attrition bias, mitigation, early termination, and clinical susceptibility bias.

Sensitivity

The probability a diagnostic test will indicate the presence of disease when the disease is actually present. In a diagnostic cohort study, the sensitivity of a test is the proportion of disease-positive results (true positives) among patients positive for the disease by the reference standard (true positives plus false negatives). Sensitivity is independent of disease prevalence and can also be calculated in diagnostic case-control studies (proportion of positive results among cases).

Specificity

The probability a diagnostic test will indicate the absence of disease when the disease is actually absent. In a diagnostic cohort study, the specificity of a test is the proportion of disease-negative results (true negatives) among patients negative for the disease by the reference standard (true negatives plus false positives). Specificity is independent of disease prevalence and can also be calculated in diagnostic case-control studies (proportion of negative results among controls).

Standardized Mean Difference (SMD)

A unitless measure of the mean difference in the value of a continuous variable between 2 groups, calculated as the mean difference divided by the pooled standard deviation for the mean in the 2 groups. The SMD is commonly referred to as “effect size,” and is frequently used in meta-analyses when similar outcomes are measured on different scales among the studies included in the analysis. As a rule of thumb, an SMD of about 0.2 is considered a small effect, 0.5 a moderate effect, and 0.8 and above is considered a strong effect. The most common versions of SMD used in clinical research are “Hedges’ *g*” and “Cohen’s *d*,” which vary slightly in their calculation of the pooled standard deviation. In DynaMed/DynaMedex evidence summaries, both are reported as just SMD.

Statistical Power

The power of a study is the ability for that study to find a statistically significant difference between 2 treatments if the difference really exists. Power depends on the number of patients in the study and the magnitude of the difference between groups. Trials are usually designed to obtain 80% power. A power calculation determines the number of participants needed to show that a prespecified difference will reach statistical significance. A study with only 30 patients might fail to show a significant difference between 2 drugs, whereas the same size difference might be significant in a study with 300 patients. When a small trial fails to find a difference but cannot exclude the possibility of a benefit that might be seen in a larger trial, the trial is said to be underpowered.

Systematic Review

A specific type of review in which researchers identify 1 or 2 specific questions, use rigorous methods to identify all available evidence, evaluate the validity of included evidence, and report their answer to the question.

Uncontrolled Trial

A single-arm interventional study in which all participants are assigned the same treatment and followed prospectively, allowing descriptive, but not comparative, assessment of efficacy and safety. Uncontrolled trials differ from case series in that they are protocol-driven, with predefined inclusion criteria and procedures. DynaMed/DynaMedex

conclusions based on uncontrolled trials always receive a level of evidence 3 due to the lack of a comparison group.

November 2023 (Version 1.02) | © 2023 DynaMed, LLC. All rights reserved.

All DynaMed and DynaMedex terms of use apply to the content within this document.