Study Design and Statistics for the Nonstatistician

Nathan S. Fox, MD

INTRODUCTION

Medical knowledge is continually evolving. As practicing clinicians, we rely on research to assess the mechanisms of disease, to learn how to best prevent, diagnose, and treat disease, and to maximize health outcomes for our patients. However, our training in understanding study design and interpreting the medical literature is usually limited to a course in medical school and ad hoc journal clubs throughout residency and beyond. Yet, we are expected to understand the implications of a given study, and how to apply the results to our own patients, both of which can be very complex processes. Some choose to "outsource" this to medical societies or hospital committees, allowing a panel of experts to review the literature and propose recommendations for the rest of us. However, having a better understanding of the medical literature and how to interpret published research remain valuable skills. They give us a better understanding of what we read, and allow us to determine if the results of a study, and the recommendations that often follow, are valuable for our own patients, either collectively, or individually. Finally, they allow us to be as informed and knowledgeable as possible when counseling our patients, improving professionalism and trust. The objective of this commentary is to present a user-friendly guide to understanding study design and statistics, designed for nonstatisticians and for doctors in clinical practice.

BASICS OF STUDY DESIGN

The Goal of Research

All research has the same goal: to find truth. With regards to any topic in medicine there exists an actual truth. Does a medication work? Will this test diagnose a condition? What is the risk of a certain outcome if I do A versus B? The issue is that none of these truths are perfectly attainable. For example, a perfectly designed clinical trial testing the efficacy of a new medication to cure cancer might, if we are fortunate, show an improvement in survival with the medication compared to placebo. However, it doesn't reveal the entire truth. Why does it work in one person but not another? What is the ideal dose for each patient? Are there certain features of the cancer that would require a different or additional medication? Are any of the side effects avoidable? Even the best study possible can only inch us a bit closer to an unreachable truth. Therefore, no single study stands alone as the answer to all questions on a topic in medicine. Additional studies are done to build on existing knowledge, examine nuances, and test new hypothesis as they are developed by others. In a sense, medical research mirrors the classic Platonic philosophy of Forms, which represent the ideal version of anything – unattainable, unreachable, and only the ideal.

Clinical Observation Compared With Research

The most basic level of inquiry that leads us towards the truth is clinical observation. This was the bulk of medical education and research for thousands of years, before study design was itself developed into a discipline. *Clinical observation* is simply noticing that certain things tend to happen. When someone gets sick, they often have a fever. People with appendicitis seem to have pain in a certain area of their right lower quadrant. If I operate this way, the patient seems to have fewer complications. Clinical observation is an important component of medical inquiry, but research is a level above clinical observation. With research, we use statistical tests to determine the likelihood of our observations being true or not.

Error

An *error* is when the conclusion we reach from our research is different from the truth. Errors are classified as type I (also known as alpha) and type II (also known as beta) errors, simply to differentiate the two ways we could differ from the truth. A type I error occurs when we conclude two things are related when they are not. If we conclude that two things are not related when they actually are, then we have a type II error. For example, if we conclude that a medication does prevent preeclampsia but the truth is it does not, we made a type I error. However, if we conclude that the medication does not prevent preeclampsia, but in truth it does, we made a type II error. There are several reasons why we could make an error, the most common of which are chance, bias, and confounding. All components of study design are meant to reduce chance and limit bias and confounding to lower the likelihood of making an error. Properly designed studies contain several elements that protect against chance, bias, and confounding, bringing the results as close to the truth as possible.

Chance

The defense mechanism against chance (also called a *random error*) is the *P* value. The *P* value is the result of statistical testing, and indicates the likelihood that the results we found were due to chance alone. A *P* value of .05 indicates that there is a 5% likelihood that the results we obtained were due to chance alone. By convention, a *P* value of .05 (5%) is often used as a cutoff for "significant" and "not significant," but this is in fact arbitrary. From a mathematical perspective, there is little difference between a *P* value of .049 and .051, although the former would indicate a "significant" finding and the latter "not significant."

Random chance is less likely to influence larger studies. Increasing the number of participants in a study or finding a larger difference between the groups will result in a lower *P* value. For example, suppose a study on women in preterm labor tested if a particular tocolytic prevented preterm birth within 2 weeks as compared to a placebo. If 10 patients were enrolled (5 in each group) and 2 of 5 in the treatment group delivered within 2 weeks, as compared to 3 of 5 in the placebo group, one probably would not be too impressed. The statistics would agree, as the comparison would be 40% (2 out of 5) vs. 60% (3 out of 5), with a *P* value of .999. Meaning, these results very likely were due to chance. Increasing the study size to 30 patients in each group with the same clinical results (40% vs. 60%), will yield a *P* value of .12, meaning there is still a 12% likelihood that the results were due to chance alone and that the results are not significant. If the study had 100 patients in each group, a 40% vs. 60% result is now significant, with a *P* value of .007, meaning there is a less that 1% likelihood those results were due to chance. More robust differences will also result in a lower *P* value. With 30 patients in each group, the *P* value was only .12 when delivery within 2 weeks occurred in 40% and 60% of the two groups. If the delivery rate within 2 weeks was 20% and 80% between the 2 groups, the *P* value is now <.001.

The 95% confidence interval is a related statistical test that can be used to test the precision of a proportion. The 95% confidence interval reports two numbers between which we are 95% confident the true number lies. Like the *P* value, the larger the number of observations, the narrower the 95% confidence interval. A simple way to think about this is with flipping a coin. A true coin (not weighted) should land on heads 50% of the time and tails 50% of the time. Suppose you were given a coin and asked to determine if it is weighted or not. If you flipped the coin 100 times and got heads 54 times and tails 46 times, you probably would not think much of it, and that would be correct. The statistical results would be that you got heads 54% of the time, but the 95% confidence interval would be 44–64%, meaning you are not 95% confident the coin will land on heads >50% of the time. However, if you flipped the coin 100 times and got heads 75% of the time, it would likely be weighted, as the confidence interval would be 66–83%, meaning you are 95% confident the coin will land on heads somewhere between 66% and 83% of the time, which would indicate the coin is weighted. If you flipped the coin a million times and got the same proportions, not only would you be certain the coin was weighted, but you would know much more precisely how weighted it is, as the 95% confidence interval would be 74.9–75.1%.

It is important to remember that the cutoff of .05 for a *P* value is arbitrary and may not always be the appropriate cutoff. Increasing the number of outcomes examined could result in a significant outcome simply due to chance alone. For example, if I compared two groups and examined 20 outcomes and use a *P* value cutoff of .05, I am allowing up to 5% likelihood that a difference in each outcome was due to chance alone. Since I am testing 20 outcomes,

that would mean I should expect at least 1/20 (5%) of the outcomes to have a *P* value of <.05. Studies that look at multiple outcomes will sometimes lower the *P* value cutoff from .05 to another number, or adjust the *P* value calculation to take into account the multiple outcomes tested (called the Bonferroni adjustment). However, not all researchers agree this is necessary.

Power and Sample Size Calculation

Power is the ability of a study to find a difference if there is one. Put another way, it is the ability to avoid a type II error. A sample size calculation is an analysis to determine how many participants are needed in a study to have a certain amount of power (usually 80% or 90%) to detect a significant difference, assuming a chance (Type I) error of less than 5% (*P* value less than .05). If a study is looking for a small difference in results (like a reduction in an outcome from 15% to 10%), more patients will be needed. As explained above in the examples regarding the *P* value, if a study does not have enough participants and a small nonsignificant difference was seen between the groups, we could have made a type II error. All well-designed prospective studies should have a sample size calculation before starting the study.

Bias

Along with chance, bias is the other reason a study could reach an error. Unlike chance, *bias* indicates that there was an actual flaw in the study design that led to incorrect findings. It is nearly impossible to eliminate bias entirely, but many aspects of study design are in place to reduce bias as much as possible. It is important to distinguish between the general connotation of "bias" and its statistical meaning. In general usage, bias implies a conscious, or possibly subconscious, favoring of one outcome over another. In statistics, bias is meant to describe a statistical difference in the actual (true) outcome vs. the outcome obtained through the methods employed in the study. So, statistical bias does not imply any conscious or subconscious effort to affect the outcome.

Selection Bias

Selection bias may occur when two groups in a study are not equal at baseline. A *P* value less than .05 in a study that compared the rates of gestational diabetes mellitus (GDM) among 100 women who took a medication and 100 who did not would seem to indicate that the medication worked. If the group of women who took the medication were all normal weight and the women who did not take the medication were all obese, it would be unclear if the lower risk of GDM was due to the medication or due to the women taking the medication having normal weight at baseline. Somehow, the women selected to receive the medication were of normal weight and the women not selected to receive the medication were obese resulting in selection bias.

Randomization is the best defense mechanism against selection bias. When a large group of participants (or anything, for that matter) are randomly divided into two groups, they should be equal at baseline. So, in the example above, had the 200 women been randomly divided into two groups of 100, and there was an equal distribution of normal weight and obese women between the groups, the lower rates of GDM in the medication would likely not be due to differences at baseline.

For studies that are not randomized, or cannot be randomized (such as retrospective studies), recognizing and adjusting for these differences at baseline is the other way to defend against selection bias, but only for differences that can be measured. These measured differences would then be considered confounding biases (see Effect–Cause and Confounding) and would be adjusted for using a regression analysis, which is a mathematical adjustment for differences at baseline. In the example above, a regression analysis would look at the rate of GDM after controlling for the differences in obesity between the groups. This would yield an adjusted risk, which, if significant, would indicate that the medication lowered the risk of GDM regardless of whether the woman was obese. Subgroup analysis is another way to control for differences at baseline. This splits the study into subgroups of women based on one of the baseline differences. In the example above, the study would compare GDM rates in women who did and did not take the medication, but do one analysis for all the obese women and another analysis for all the nonobese women. This is a good option if the study size is large enough. Otherwise, each subgroup analysis might not have enough participants to reach statistical significance (see the section on Chance, above). Matching is a third way to control for differences at baseline. This method selects specific control patients who are matched to the case patients for

a certain characteristic, such as age or weight. In the example above, for each woman who took the medication, a control patient would be selected who had a similar weight to the case patient, thus matching by maternal weight. This method would eliminate the influence of maternal weight on the outcome, but would also limit the ability to ascertain the effect of maternal weight on the outcome.

Information Bias

Information bias is when the quality of information obtained from one group is different from the quality of information obtained from another group. This can be seen in prospective as well as retrospective studies and can manifest in several ways. In a prospective study, a participant's knowledge that she is receiving a certain treatment might affect how likely she is to report an outcome or side effect. A researcher's knowledge of a treatment might also effect the likelihood of identifying an outcome or side effect in the participant. Any differences in outcomes or side effects seen between the groups could be due to the different quality of information. In retrospective studies, since the information has already been obtained, it is possible that patients with certain conditions, or receiving certain treatments had more data collected and recorded in the medical record. It is possible that the controls simply do not have certain outcomes noted in the medical record because nobody thought to assess or record them. For example, women taking a medication may be asked many more times if they have certain side effects than women not taking a medication. If more women in the medication group reported a certain side effect, it is unclear if it was due to the medication itself, or simply because they were asked and the other group was not.

Recall bias can occur in survey studies when patients are asked to recall past events. Patients who experienced certain outcomes might be more likely to recall a risk factor than women who did not have the outcome. A typical example of this is with teratology studies. If a researcher asks a group of women to list all the medications they took in pregnancy, the women who delivered babies with birth defects might be more diligent about listing every medication they took the entire pregnancy, as compared to women who delivered babies without birth defects. This could lead to an incorrect conclusion that a certain medication is associated with the birth defect.

There are several defense mechanisms against information bias. Blinding the participants and researchers in a prospective study as to which group they are in should minimize any differences in the quality of information collected between the two groups. This is one reason why medication studies often will have a placebo given to the control group (the other reason being that simply taking any medication might actually have a biological effect on the participant, also known as the *placebo effect*). If blinding is not possible or feasible, using a prespecified set of outcomes that are as objective as possible will help to reduce bias. For example, a study with cesarean delivery rates as an outcome is less prone to information bias, as that outcome is objective. However, a study with pain scores or estimated blood loss as outcomes are more prone to information bias.

In retrospective studies, it is more difficult to defend against information bias. If it is a survey study, it is imperative not to inform the participant what the hypothesis of the study is. If that is not possible, it is simply a limitation of these types of studies.

Treatment Bias

Treatment bias is related to information bias. Instead of the quality of information being different between the groups, with treatment bias (co-intervention bias), the actual clinical treatment is different between the groups (in addition to the treatment being studied). If a study compared two medications used for induction of labor and the outcome was time to delivery, it is possible that the treating doctors would manage labor differently if they knew which induction method the patient received. Any differences seen between the groups with regards to time to delivery could be due to the different labor management styles, as opposed to the different induction methods used. Similarly, the decision to do things like perform a cesarean delivery, give a blood transfusion, or administer antibiotics, are all under the control of the doctor and could be subject to treatment bias.

The defense mechanisms against treatment bias are similar to those for information bias. Blinding the researchers, if possible, is the best mechanism. If this is not possible or feasible, it is important that a prospective study prespecify how patients are to be managed in each group. Blinded studies need less prespecificed rules because the researchers don't know who is in each group; unblinded studies need more prescecified rules and treatment protocols to keep treatment as similar as possible between the

groups. If the study is retrospective, the researchers should try to consider all the possible differences in management and account for them in a regression analysis, or at least recognize them as limitations in their study.

Effect–Cause and Confounding

If a study concludes that A and B are related, there are several possible explanations. The first is that A causes B. Two other options are that B causes A (effect–cause), or that there is some third variable, C, which confounds the relationship between A and B (Figure 1).

If a study concluded that emergency cesarean deliveries (A) and low newborn Apgar scores (B) were associated, it would be incorrect to conclude that A causes B, that emergency cesarean deliveries cause low Apgar scores. It is more likely that B causes A, that babies who have low Apgar scores also had fetal heart rate monitoring abnormalities, which led to the emergency cesarean delivery. This is an example of an effect–cause misinterpretation leading to a type I error. The defense mechanism against this is usually performing a prospective randomized study. Certain studies, such as the example of emergency cesarean deliveries and Apgar scores, do not lend themselves to that study design. In this case, the analysis of the data must be thoughtful and consider effect–cause as a possibility in interpreting the results.

Confounding is when an unreported third variable (C) is truly the cause for the relationship between A and B. This can happen in two different ways (see Figure 1). The first is when C causes both A and B. If a study found that oxytocin use (A) and cesarean delivery (B) were related, it would be incorrect to conclude that A causes B, that oxytocin use causes cesarean delivery. It is likely that there is a confounding variable, protracted labor (C), which causes both an increased use of oxytocin (A) as well as an increased risk of cesarean delivery (B). If a study concluded that antibiotics in labor (A) were associated with neonatal brain injury (B), it is likely that maternal infection (C) is a confounding variable that causes both the use of antibiotics in labor (A), as well as neonatal brain injury (B). When the groups are not equal at baseline (selection bias), this is a classic form of confounding. The best defense against this type of confounding is prospective study design with randomization.

The other model of confounding is when the third variable (C) is an intermediary between A and B (see Figure 1). If a study found that breech presentation (A) was associated with postpartum endometritis (B), it is most likely that the reason this is true is because breech presentation leads to cesarean delivery (C), which is associated with an increased risk of endometritis. It would not be inaccurate to report that breech presentation leads to endometritis, but it would not be precise, and would in fact be misleading because it is the cesarean delivery, not the breech presentation itself, that increases the risk of endometritis. Since it is not technically an error to conclude that breech presentation causes endometritis, there is no specific defense mechanism against this in study design, aside from proper framing of the hypothesis and careful interpretation of the results. A careful analysis of this example would easily discover that breech presentation itself does not lead to endometritis because the hypothesis is not plausible and because cesarean delivery is an obvious clinical intermediary.

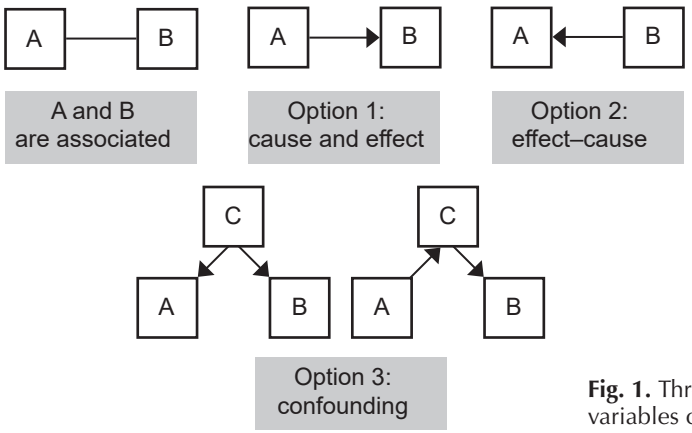

Fig. 1. Three options for how two variables can be associated.

Unless confounding is properly addressed, for any study that shows a relationship between A and B, it is not appropriate to automatically conclude that A causes B, or that A increases the risk of B. Rather, the appropriate conclusion would be that A and B are associated.

A summary of the causes for error in research, as well as the defense mechanisms used to prevent error can be found in Table 1.

PUTTING A STUDY INTO CONTEXT

Even if a study is well-designed, limits bias, and has statistically significant results, it is important for the reader to consider whether the results are clinically useful.

Does the study apply to every population?

If a study is conducted in a population of women with different demographics, the results may not be applicable to other women. Or, if it was conducted in a different time period when other medical care was different, it may not apply today.

Relative Risk Compared With Absolute Risk

It is important to consider if the differences in outcomes are clinically meaningful or just statistically meaningful. For example, if a study evaluating treatment of postpartum hemorrhage showed a statistically significant decrease of blood loss, but the absolute decrease was only 50cc, it might not be clinically relevant. Similarly, consider a study that concludes that having a family history of preterm birth is associated with a 30% increased risk of preterm birth. Since the risk of preterm birth is only 10% at baseline, that would mean that a woman with a family history of preterm birth only had her risk increase by 3%, from 10% to 13%. In this study, the relative risk is 30%, but the absolute risk is only 3%. Although the relative risk may be statistically significant, the relative risk is not likely to be clinically significant. The interpretation of what is and is not clinically significant depends on the details of the condition, treatment, or side effects.

STUDY TYPES

Cohort Study

A *cohort study* compares two or more groups that are differentiated by an exposure and measures the differences in an outcome (or outcomes).

The exposure that differentiates the groups can be something inherent to the participant (age, weight, smoking status), or can be something introduced in the study (medication, operation, intervention).

Features of a cohort study

- It can be done prospectively or retrospectively.

Table 1. Causes for Error in Research

Reason for Error	Defense Mechanism in Study Design
Chance	*P* value 95% confidence interval
Selection bias	Randomization
Information bias	Prospective study with prespecified data points Objective, as oppposed to subjective, outcomes Blinding
Treatment bias	Blinding the researcher In unblinded studies, prespecify treatment protocols for each group
Effect-cause	Prospective, randomized study design Thoughtful analysis of retrospective or observational data to consider effect-cause as a possibility when interpreting the results
Confounding	Prospective study design Regression analysis Subgroup analysis Matching Randomization

• If prospective, the study can be randomized, blinded, or placebo-controlled, based on the specific exposure chosen (for example, one cannot randomize women into different maternal age groups).

• The best form of a cohort study is a randomized controlled trial (RCT). However, many cohort studies are observational, ie, not randomized.

• A primary outcome is chosen, and this is what is used to perform the sample size calculation. Secondary outcomes can also be examined.

• The study will yield a relative risk (RR), which is the risk of the outcome in one group compared to the other. A RR of 2.0 means one group is twice as likely as the other group to have the outcome. This can also be reported as a RR of 0.5, meaning the risk in the second group is one half the risk in the first group. Those two RR's are basically the same, but depend on which group you choose as the reference group. The RR will be followed by a *P* value or a 95% confidence interval, which will determine if the differences seen are statistically significant. If a group had a 30% incidence of preterm birth and the other group had a 10% incidence, the results might look something like: 30% vs. 10%, RR 3.0, 95% CI 1.8–5.3, *P*=.002), which would indicate a significant difference.

Things to consider when reading a cohort study

• If the study was not randomized, were the groups equal at baseline? If not, was this addressed by the researchers with further analyses?

• If the study was not blinded, were the groups managed the same, aside from the single exposure being studied?

• Was the right outcome chosen? Meaning, was the outcome clinically meaningful?

• Was the quality of information collected between the two groups similar?

• Were there any potential confounding variables not addressed by the researchers?

• Are the outcomes between the groups significantly different? If so, is the absolute risk clinically meaningful?

• If no difference was found between the groups, were there enough patients to be sure a type II error was not made (was a power analysis done)?

• Do the findings apply to your own patient(s)?

Case-Control Study

A *case-control study* compares two groups differentiated by an outcome and examines if risk factors for that outcome differ between the groups.

In a case-control study, cases are those with the outcome and controls are either matched to cases or are all participants without the outcome. Therefore, a case-control study is like a mirror image of a cohort study. In a cohort study, the groups are differentiated by the exposure and differences in outcomes are ascertained. In a case-control study, the groups are differentiated by the outcome and differences in exposures are ascertained. For example, a case-control study for preterm birth might look at a group of women with preterm birth <37 weeks of gestation (cases) and compare them to women who delivered ≥37 weeks of gestation (controls). Then differences in exposures would be compared between the two groups, such as age, bleeding during pregnancy, or activity during pregnancy. These data can be ascertained through medical records, patient interviews, stored samples, etc. A case-control study calculates an odds ratio (OR). Although the OR is frequently similar to the RR (especially with larger studies), they are different. An OR calculates the odds of one group having an exposure as compared to another group (one odds divided by another odds, which is why it is called an odds ratio). The statistics will be reported as an OR with a 95% confidence interval, or as an adjusted OR and 95% confidence interval, if the analysis took into account several exposures and did a regression analysis.

Features of a case-control study

• It can only be done retrospectively.

• It is usually done to look at risk factors for a rare outcome, but the outcome does not need to be rare.

• There is always a risk of having unmeasured confounding variables that are not included in the analysis.

• There is potential for bias given the retrospective nature.

Things to consider when reading a case-control study

• Was the control group selected independent of exposure?

• Were there important exposures not included in

the study?

- If the exposure information was obtained from patient interview, was there a possibility of recall bias or information bias?
- Case-control studies can only determine associations, not causation.

Other Study Types

Meta-analysis

A *meta-analysis* combines the data from several studies and performs a statistical analysis on the pooled results. It is often done either to reconcile several studies with differing findings, or to add smaller studies together to increase the sample size and therefore the power. A meta-analysis is limited by the quality of the studies included, so all meta-analyses should report an analysis on the quality of each study included. They are also prone to publication bias, which is when only certain studies on a topic are published and available for meta-analysis. Finally, they are prone to selection bias if the authors choose to include certain studies and exclude others.

Observational Descriptive Studies

An *observational study* simply reports the prevalence of a condition for a large group of patients, but does not compare two groups. A study that reports the incidence of blood transfusion in women undergoing a first, second, third, and fourth cesarean delivery, but does not compare those results, would be observational. Observational studies are an important part of scientific inquiry as they are often the studies from which research questions are proposed, which lead to more robust studies.

Case reports (one patient) and *case series* (more than one patient) report a novel or interesting clinical presentation, treatment, or outcome. They are not considered research but are an important part of the investigative process as they are usually the basis for further inquiry.

Reporting Guidelines

In order to promote transparency, accuracy, and timeliness of the reporting of research studies, there have been several international initiatives to standardize and improve the way authors report their research. These include CONSORT for randomized trials, STROBE for observational studies, and PRISMA for meta-analyses, as well as others. Many scientific journals require authors to submit evidence they have followed these guidelines. More information can be found at http://www.equator-network.org/toolkits/.

SCREENING TESTS

Studies designed to test the characteristics of screening test are a separate type of research.

In general, there are two reasons to perform a screening test:

1. To select a subgroup of patients from a larger population to undergo a diagnostic test, because the diagnostic test itself is either expensive, painful, or has risk associated with it. Examples of this kind of test are the glucose challenge test, aneuploidy screening, mammography, and cervical cancer screening with a Pap test and human papillomavirus testing.
2. To find a subgroup of patients at increased risk of a certain condition, either to inform them of risk, or to try to intervene to prevent the outcome. An example of this kind of screening is cervical length screening in pregnancy.

Screening characteristics

Each screening test has its own set of test characteristics, which help quantify how well the test predicts the outcome. Screening tests can be characterized several ways, which are listed below, along with their meaning using the example of a cervical length ≤25 mm at 20 weeks of gestation as a predictor for preterm birth.

a. *Sensitivity:* The percentage of people with the outcome who the screening test will call positive. For example, of all the women who deliver preterm, how many will have a cervical length ≤25 mm at 20 weeks of gestation. A sensitivity of 80% means that 80% of women who deliver preterm will have a cervical length ≤25 mm at 20 weeks of gestation. Another way to look at it is a sensitivity of 80% would than have a 20% false-negative rate, meaning 20% of women who will ultimately deliver preterm will be "missed" by a cervical length screen.

b. *Specificity:* The percentage of people without the outcome who the screening test will call negative. For example, of all the women with term births, how many will have a cervical length >25 mm at 20 weeks of gestation. A specificity of 80% would mean

that 80% of women who deliver at term will have a cervical length >25 mm at 20 weeks of gestation. Another way to look at it is a specificity of 80% would than have a 20% false-positive rate, meaning 20% of women who will deliver at term will have a positive cervical length screen.

c. *Positive predictive value (PPV):* The percentage of people who test positive who will have the outcome. For example, of all the women with a cervical length ≤25 mm, how many will deliver preterm? A PPV of 80% would mean that 80% of women with a cervical length ≤25 mm at 20 weeks of gestation will deliver preterm.

d. *Negative predictive value (NPV):* The percentage of people who test negative who will not have the outcome. For example, of all the women with a cervical length >25 mm, how many will not deliver preterm. A NPV of 80% would mean that 80% of women with a cervical length >25 mm will not deliver preterm.

e. *Positive likelihood ratio (+LR):* The probability that someone with the outcome will have a positive test, as compared to someone without the outcome. It is also defined as the sensitivity/(1-specificity). The results are listed as a number, such as 2.0, which would mean that someone who delivers preterm is twice as likely as someone who delivers at term to have a cervical length ≤25 mm at 20 weeks of gestation. It also means that someone with a cervical lenth ≤25 mm is now twice as likely to deliver preterm.

f. *Negative likelihood ratio (–LR):* The probability that someone without the outcome will have a negative test, as compared to someone with the outcome. It is also defined as (1-sensitivity)/specificity. The results are listed as a number, such as 0.5, which would mean that someone who delivers at term is one half as likely as someone who delivers preterm to have a cervical length ≤25 mm at 20 weeks of gestation. It also means that someone with a cervical length >25 mm is now half as likely to deliver preterm.

Specifics about screening test characteristics

Sensitivity and specificity are features of the test itself and do not vary based on the population being studied. The cervical length screen for preterm birth, for example, should have the same sensitivity and specificity in two populations of 1,000 similar singletons, regardless of how many women in each group of 1,000 actually deliver preterm. The same is true for +LR and –LR as they are based on the sensitivity and specificity. However, the PPV and NPV depend on the incidence of the outcome in the population. For example, suppose cervical length is used to screen two populations of 1,000 women for preterm birth (Group A and Group B). In Group A, 300 (30%) women deliver preterm, and in Group B only 50 (5%) women deliver preterm. The sensitivity and specificity will be the same for each group, but the PPV will be higher in Group A than Group B, and the NPV will be higher in Group B than Group A. This is because a positive test in Group A is more likely to lead to preterm birth because more women in that group had preterm birth. It is important to realize this when interpreting PPV and NPV. For example, if a study concluded that a 20-week anatomy ultrasonogram had a 98% NPV for aneuploidy, it would initially appear to be a very impressive screening test: a woman with a normal ultrasonogram has a 2% or less chance of having a baby with aneuploidy. However, since the incidence of aneuploidy at 20 weeks of gestation is likely to be less than 2% anyway, the test is really no better than doing nothing. However, if the ultrasonogram had a 98% NPV for aneuploidy among women with an abnormal serum screen for aneuploidy, it might be more valuable.

The characteristics of a screening tests can manipulated based on the definition of an abnormal screening test. Many screening tests yield a continuous result (meaning, not yes or no, but rather a number or value) and a decision needs to be made which value will be the cutoff for normal and abnormal. For example, the choice to use a cutoff of 25 mm for the cervical length screen could be changed and it would affect the test characteristics. If the cutoff was lowered to 10 mm, now only women with a really short cervical length are called abnormal. This would lower the number of people we will identify, but people who test abnormal will now have a higher risk of preterm birth. This change to 10 mm will therefore lower the sensitivity and NPV, but increase the specificity and PPV.

For this reason, sensitivity and specificity are rarely useful in isolation. One could manipulate a screening test to have 100% sensitivity, but it usually comes at the expense of a very low specificity (and vice versa). Ultimately, the decision of what cutoff to use is based on the severity of the outcome and the expense and pain of any treatments or secondary tests. Frequently, studies will use a receiver operator characteristic curve (ROC curve) to determine the optimal cutoff.

The general considerations for choosing a screening test are shown in Figure 2 and use the glucose challenge test as an example. Clinically, a glucose challenge test value somewhere between 130–140 mg/dL is often used to screen for GDM. For those who have an abnormal (elevated) glucose challenge test, they proceed to take a second and longer 3-hour test. Using a cutoff of 130 mg/dL would increase the detection of GDM, as there will be fewer false-negative glucose challenge tests (fewer "misses"), but that would come at the expense of having more people without GDM testing positive (more false positives, or "scares"). On the other hand, using a glucose challenge test cutoff of 140 mg/dL will do the opposite. There will be less "scares" but more "misses." Ultimately, there is no exact right answer for what cutoff to use for a screening test, as it depends on the specific condition being screened for and the specifics of the population being screened.

FINAL THOUGHTS

Understanding research methodology and study design is an important part of reading the medical literature and practicing medicine. Experience is very important in medical practice, but humans have selective memory, and over time, we have found that many of our observation-based assumptions were incorrect. True, many medical questions do not have solid evidence-based research upon which to rely and decisions must be made based on judgement alone. However, it is imperative to understand when a decision is based on judgement alone and to be open to changing practice if well-designed research shows another management is superior.

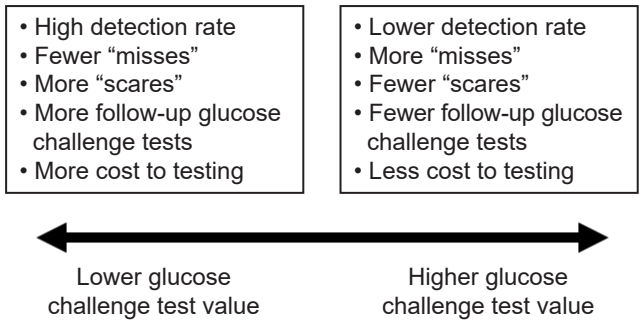

Fig. 2. Considerations for choosing a glucose challenge test value as a cutoff to screen for gestational diabetes.

