Statistical Thinking: Clinical Versus Statistical Significance

Robert A. Calder, MD, MS; Jayshil J. Patel MD

The Medical College of Wisconsin's FUSION curriculum includes 7 threads that are woven through the different phases of undergraduate medical education.

Under the premise that an accurate and timely diagnosis is the cornerstone of medicine and that evidence informs clinical decision-making, one of the seven threads – the "Critical Thinking in Medicine" thread – aims to unravel the diagnostic process and describe the fundamental components of evidence-based medicine. In the first of a series of articles on this thread, we will outline the statistical thinking needed to differentiate clinical and statistical significance.

In a subsequent article, we will introduce a story-like format to describe the statistical thinking needed to solve common problems encountered in clinical medicine, without memorizing a formula and without a calculator. For example, suppose a middle-aged woman undergoes routine screening mammography and is found to have a positive test. Assuming that the prevalence of breast cancer in such women is 0.8%, that mammography is 90% sensitive for detecting breast cancer, and that there is a 7% false-positive probability, what is the probability that the positive test represents breast cancer? The most common answer physicians give is "90%." But, as you will see, it closer to 10%.[1]

Statistical thinking is at the heart of biostatistics, which is perhaps the most frequently used basic science. In his excellent brief text, *High-Yield Biostatistics, Epidemiology and Public Health,* Anthony Glaser states, "There is perhaps no other area in USMLE Step 1 from which knowledge will be used every day by every physician, no matter what specialty they are in, and no matter what setting they are practicing in."[2]

Thinking vs Statistics

Perhaps the first step in learning statistical thinking is to realize that the numbers, per se, are far less important than clinical judgment. In other words, thinking in context is more important than interpreting the statistics alone. Whether some difference between 2 treatments is statistically significant is a mathematical, probabilistic judgment. However, deciding whether any difference between treatments is meaningful – and important to patients – is a clinical judgement that does not involve mathematics. For example, suppose a new treatment for dementia becomes available and is supported by a large scientific study showing a highly statistically significant difference between subjects who received the treatment versus placebo. Would this be a useful treatment? Should patients and other payors be willing to pay for it? The usefulness of such a therapy should depend far more on the patient's values and preferences and how large and meaningful the difference is between those who received the treatment versus the placebo rather than the statistical significance of the difference. A small but statistically significant change in some cognitive performance scale that is difficult to interpret may not be clinically meaningful regardless of its statistical significance.

Statistical significance is measured by the *P* value. The *P* value tells us how likely the difference between treatments – or even more extreme differences – could have occurred due to chance alone if there is truly no difference between the treatments. The assumption of "no difference" between the treatments is commonly called the "null hypothesis." Whether a difference between treatments is statistically significant can certainly be helpful in determining whether the treatment may be useful, but there are other important considerations.

Testing a Null Hypothesis

To highlight these considerations, we will describe a very simple application of testing a null hypothesis and determining statistical significance. Suppose a friend gives you a coin that she says "lands on heads far more often than tails" (analogous to her saying that she has discovered a treatment that is more beneficial than another). How can we test such claims? The first step is to state a null hypothesis. In this case, our null hypothesis would be

Author Affiliations: Medical College of Wisconsin, Milwaukee, Wisconsin (Calder); Division of Pulmonary and Critical Care Medicine, Medical College of Wisconsin, Milwaukee, Wisconsin (Patel).

Corresponding Author: Robert A. Calder MD, Adjunct Assistant Professor, Medical College of Wisconsin, Milwaukee, WI; email rcalder@mcw.edu.

to assume that the coin is "fair" – that is, we will assume that it is just as likely (50-50 chance) to land on heads as tails. Then, we proceed to do an experiment, we flip the coin some number of times and see how it lands. If it lands on heads far more often than we would expect by chance alone, at some point we might be willing to "reject" the null hypothesis that the coin is fair (and accept an alternative hypothesis, eg, that the coin is not fair). However, how often does the coin need to land on heads for us to reject the null hypothesis of no difference? Let's assume that the coin lands on heads 4 times in a row. Is that good enough to assume that it is not "fair"? If the coin is fair, the probability of it landing on heads 4 times in a row is simply the product of the probability of it landing on heads each time which is: ½ x ½ x ½ x ½ which is 1/16 or 0.0625. Is that good enough? What if the coin landed on heads 9 times in a row? If the coin is fair, the probability of that is (1/2)^9 which is 0.00195 or about 2 chances in 1000. Are you ready to declare that the coin is not fair? What if your life depended on this question? Would this be enough evidence to bet your life on the next flip of the coin? Most people would probably want stronger evidence before betting their life on this coin.

The need to assess the strength of the evidence is the first lesson in understanding *P* values. Again, the *P* value is the probability of getting these or even more extreme results if there is no difference in the treatments being considered (heads or tails in this example). The P value that you select to allow you to declare a difference "statistically significant" should depend on the clinical situation, not just whether it is above or below the "traditional" level of statistical significance, commonly 0.05. A significance level of 0.05 means that there is a 1 in 20 chance of seeing a difference that large or even larger, if there is truly no difference between the treatments. That may be good enough if the clinical question being considered is "routine." You may even be satisfied with a *P* value of 0.10 if you are considering a treatment for a life-threatening disease for which no treatment currently exists. On the other hand, if you are considering suggesting a potentially toxic therapy with many possible side-effects to a patient – perhaps to treat a life-threating cancer – you may demand a level of statistical significance far greater than 1 in 20 (0.05) before you are willing to declare the difference statistically significant and to suggest this treatment for a patient. In essence, we consider that the level selected for statistical significance should depend on the clinical situation and not simply the "traditional level" of 0.05.

Box: Key Terms

Null Hypothesis: A statement proposing that there is no difference between the treatments being considered.

***P* value:** The probability that results obtained or more extreme results could have occurred due to chance alone assuming that the null hypothesis is true.

Type I error: Rejecting the null hypothesis when it is true. The null hypothesis is really true, there is no difference but you erroneously rejected that hypothesis. A false alarm.

Type II error: Failing to reject the null hypothesis when it is wrong. There is a difference, but you failed to reject the null hypothesis of no difference. You missed the boat.

Alpha (α): The probability of a type I error, a false alarm (mnemonic: alpha and alarm both start with "a").

Beta (β): The probability of a type II error, you missed the boat (beta and boat both start with "b").

Power: 1 – β; 1 minus the probability of missing the boat is the probability that you got on the boat. This is the ability of a study to designate a difference of some size between treatments as statistically significant at some given level.

Number Needed to Treat (NNT): The number of subjects who would need to be treated over some time period in order to prevent one bad event or cause one "good" event.

What is the Origin of the Traditional *P value* Level of 0.05?

The traditional significance level of 0.05 was originally suggested by R. A. Fisher, who performed agricultural experiments in England starting in the 1920s. In 1925 he suggested that "We shall not often be led astray if we draw a conventional line at 0.05…"[3] He certainly did not imply that 0.05 should be considered a Rubicon that establishes a result as important regardless of the clinical situation. It is unfortunate that the word "significant" is associated with any given number since this word implies a level of importance that may not be appropriate.

Testing Our Null Hypothesis

Returning to our coin-flipping experiment, suppose we state our null hypothesis that the coin is "fair" (ie, no difference between heads and tails) and pose an alternative hypothesis that the coin is not fair. We then specify a significance level of less than 0.0001 to denote statistical significance (because we want to be really sure). Then we flip the coin 100 times and find that it lands on heads a total of 75 times out of 100 (in any order, not necessarily all in a row, of course). What is the probability of this happening, if the null hypothesis is true? That probability is about 0.000000287, or about 3 chances in 10 million; we therefore reject our null hypothesis and accept the alternative hypothesis that the coin is not fair. Recall that the probability of 9 heads in a row was about 2 in 1000. With 100 flips, it is far less likely to get a total of 75 heads, even though many people would think 9 heads in a row seems perhaps more impressive. This example demonstrates why computing *P* values is valuable: we humans are not good at judging probabilities using our "gut instincts." For example, if you flip what you think is a fair coin and it lands on heads 4 times in a row, you are likely to think you are "due" for a tails. Well, you aren't. That's the kind of thinking that keeps casinos open (the Gambler's Fallacy). The probability of getting a tail on the fifth flip is 50%, just as it is for each of the other flips. Any "imbalance" of flips early on (for example 4 heads in a row) is not corrected but rather "diluted" as the flipping continues.[4] This is a very common misconception and is one of the reasons that computing *P* values is worthwhile. We humans have many misconceptions about probability.

Courtroom Analogy

Hypothesis testing is analogous to what happens (or should happen) in a courtroom. Our null hypothesis is analogous to assuming that the defendant is innocent. We then see evidence presented in court. If we are trying someone accused of robbing a bank and several witnesses identify the defendant as the guilty party, and the police find the money from the bank in his apartment, the jury will prob-

ably determine that such evidence would be very, very unusual if the defendant were indeed innocent. So they then reject the idea that he is innocent and declare him guilty.

Hypothesis Testing Errors

As in the courtroom, there is the possibility of convicting the innocent and letting the guilty go free. We "convict the innocent" when we claim that some result is statistically significant when it really is not. The probability of this happening is the level that we set for determining statistical significance, eg, 0.05. This is also denoted as "alpha" (α). So, the probability of claiming a difference when there is really no difference is alpha, also called a "Type I error." There are many confusing ways of stating a Type I error, but we prefer to state it as a false alarm. You claimed there was difference, but there really wasn't (false alarm).

On the other hand, it is also possible to fail to find a difference that is truly there, analogous to letting the guilty go free. The robber committed the crime, but the evidence wasn't sufficient to reject the assumption of innocence. If a study is small, there is a greater chance of failing to find an important difference between treatments; the probability of this is "beta" (β). This is also called a "Type II error," but we prefer to think of it as missing the boat. There was a difference there but you missed it, because your study wasn't powerful enough to pick it up. Power is the ability of a study to find a difference if it is present. So, given that beta is the probability of "missing the boat" then 1 minus beta equals the probability that you got on the boat –that is the "power" of the study.

Large Study, Small Study

Large studies are likely to find small differences to be statistically significant. For example, suppose that 545 out of 1000 subjects receiving "treatment A" improved, whereas only 500 out of 1000 subjects improved with "treatment B." This small difference (54.5% vs 50.0%) is statistically significant at the 0.05 level ($P=0.0488$, chi-square test). On the other hand, suppose that 9 out of 10 (90%) improved with Treatment A versus only 5 out of 10 (50%) with treatment B in a small study. With this small study, a large proportionate difference (40%) was not statistically significant at the 0.05 level ($P=0.1409$, chi-square test).

Don't Believe All Headlines

When the headline of a study includes the words "no significant difference," always ask yourself, how big was that study and what was its power to pick up a clinically meaningful difference? On the other hand, when a large study touts that treatment A is significantly better than treatment B, always ask yourself, how big is the difference? Would this size difference be clinically meaningful to me and my patients? How many patients would I have to treat in order to see 1 patient benefit? (Number needed to treat [NNT] is the subject of a future column.)

These considerations highlight the difference between statistical significance, which is a mathematical benchmark, and clinical significance, which is determined by the patient, the clinician, and the clinical situation.

Acceptance Region

One final point about the *P* value is worth emphasizing. The *P* value is a dichotomous indicator, ie, it either is or is not statistically significant. Either the study found the *P* value was below the significance threshold or it was above it. The "acceptance region" provides much more information. The acceptance region provides a range of values that, if the null hypothesis is true, we could expect our data to fall. For example, if we expect 50 heads in 100 flips of the coin, probability theory (the subject of another future column) tells us that there is a 95% chance that we will get between 40 and 60 heads. If we get fewer than 40 or more than 60, that is unusual and, in the long run, will only happen 5% of the time with a fair coin. This provides much more information than just an "up" or "down" decision using the *P* value. In a future column, we will discuss how to calculate acceptance regions and a related concept called the confidence intervals. At this point, it is worth understanding that both acceptance regions and confidence intervals provide more information than the *P* value because they provide a range within which you would expect your data to reside.

Rethinking The Evidence

One of the most exciting aspects of medical science is having our longstanding beliefs and practices shown to be wrong with new evidence. The most important thing that evidence should do is to cause us to rethink what we believe. Our intuition can fool us into making erroneous conclusions. New evidence, while potentially upsetting, can and should challenge our preexisting beliefs. Fortunately, statistical thinking allows us to evaluate new evidence and clearly discuss it with patients in a way that incorporates their unique values and preferences.

Conclusion

In summary, we use experiments (clinical trials) to decide which treatment is better starting with the assumption (the null hypothesis) that both are the same. If the experiment provides results that would be very unlikely if the treatments were the same, we can then "reject" the idea that the treatments are the same with some given level of confidence (alpha). Most important is whether a difference between treatments is clinically significant rather than just statistically significant. Clinical significance is determined by using clinical judgment and talking with the patient and identifying their values and preferences when applying the results of evidence. This principle of evidence-based medicine is far more important than simply achieving a statistical benchmark. With a large enough study, trivial differences can be "made" statistically significant. On the other hand, small, underpowered studies can fail to find important differences. It is critical to know what you would consider a clinically important difference and what the study power is to find that difference.

In the next column in this series, we'll discuss risk: absolute risk, relative risk, and the "number needed to treat." This is perhaps the most important skill needed to communicate medical information in a manner that informs rather than confuses patients.

Financial Disclosures: None declared.

Funding/Support: None declared.

REFERENCES

1. Gigerenzer G. Calculated Risks, How To Know When Numbers Deceive You. Simon and Shuster; 2002

2. Glaser AN. High-Yield Biostatistics, High Epidemiology, and Public Health. 4th ed. Wolters Kluwer; 2014

3. Fisher RA. Statistical Methods for Research Workers. Oliver and Boyd, 1950:80.

4. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science, 1974;185(4157): 1124-1131. doi:10.1126/science.185.4157.1124

advancing the art & science of medicine in the midwest

WMJ

WMJ (ISSN 1098-1861) is published through a collaboration between The Medical College of Wisconsin and The University of Wisconsin School of Medicine and Public Health. The mission of *WMJ* is to provide an opportunity to publish original research, case reports, review articles, and essays about current medical and public health issues.

Visit www.wmjonline.org to learn more.