POSNA Evidence-Based Practice Committee's Primer on Statistics

Sara Davis, MPH[1]; Neeraj M. Patel, MD, MPH, MBS[2]; Ifeoma Inneh, MPH, MBA[3]; Raymond Guo, BA[4]; R. Justin Mistovich, MD[5]; Tracey Bastrom, MA[6]; Scott D. McKay, MD[3]; POSNA Evidence-Based Practice Committee

[1]Texas Children's Hospital, Houston, TX; [2]Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL; [3]Baylor College of Medicine & Texas Children's Hospital, Houston, TX; [4]Baylor College of Medicine, Houston, TX; [5]Rainbow Babies and Children's Hospital, Cleveland, OH; [6]Rady Children's Hospital, San Diego, CA

Introduction

Statistics describe the major findings of every research paper we read and are often the basis of the decisions we make daily taking care of patients in the clinic and the operating room. Because statistics are relatively quantitative, orthopaedists and other engineering types can wrap their heads around differences in numbers. Yet statistics don't tell the whole story, and as some have quoted, "… not everything that can be counted counts, and not everything that counts can be counted." Stated another way, findings that are statistically significant may not be clinically important. For those of us who have spent greater than 10 years in clinical practice, the interval from freshman biostatistics is increasing, and how we critically examine literature may have become less rigorous. The POSNA Evidence-Based Practice Committee has provided a quick refresher on the important aspects of critical literature review, and this article serves as a review of common statistical terms in a "case-based" format. We will explain **standard deviation, p-value, number needed to treat, confidence interval, sensitivity, specificity, and negative and positive predictive values.**

Case Example

Dr. Bone was recently "informed" by the newly minted administrative clinical dyad (Mr. H.I. Energy) that her clinic patient satisfaction scores have been decreasing. Dr. Bone pointed out that most of her patients like

Figure 1. *On average, this boy has a satisfaction score of 8.6 with his clinic visit.*

stickers that state, "I was brave today" or that depicted recent Disney characters frozen in time; thus, she posited that perhaps a nine-year-old's satisfaction may not be clinically important? Despite this, her energetic administrative associate encouraged her to consider a QI project that may improve patient satisfaction in her pediatric orthopaedic clinic. With a grant from Tootsie Roll Industries (Chicago, IL), and after obtaining institutional review board (IRB) approval, Dr. Bone conducted a prospective study in which children were blindly assigned to either (a) receive a randomly selected lollipop after their appointment or (b) not receive a lollipop. She then collected data on patient satisfaction, with 10 being the best score and one being the worst. Upon conclusion of

the study, 1000 patients were included in analysis, of which 500 received a lollipop and 500 did not. She reported to her colleague that the mean satisfaction of the lollipop group was 8.6 (standard deviation 1.2). Mr. Energy was uncertain if this was useful information.

The standard deviation (SD) describes how widely (or narrowly) data is spread in relation to the mean of a continuous variable. If the SD is low, then the data are tightly grouped around the mean. If the SD is high, then there is a wide range of values around the mean. The SD is most easily interpreted if the data of interest is normally distributed. In this case, the data are symmetrically distributed around the mean, and the mean, median, and mode are all equal. In such a scenario, 68% of all values are within 1 SD of the mean, 95% are within 2 SD, and 99.7% are within 3 SD. If the data in the above example is normally distributed, this would mean that 68% of all subjects had a satisfaction score between 7.4 and 9.8 [the mean of 8.6 ± 1.2]. If the data is not normally distributed, additional descriptors may aid in qualifying it (i.e., interquartile range). However, the SD can still be thought of as an indicator of variation or dispersion.

Dr. Bone then revealed to Mr. Energy that the non-lollipop group had a mean satisfaction score of 5.3 with a standard deviation of 1.4 (compared to 8.6±1.2 for the lollipop group). After ensuring that the data was normally distributed, she conducted an independent samples t-test to compare the means of the two study groups. This statistical test is best for continuous variables that are normally distributed, such as age or satisfaction scores in the current study. The p-value was less than 0.0001. She also wondered if the proportion of males in the lollipop group (54%) was different from the proportion in the non-lollipop group (49%). Because these are non-continuous variables, she conducted a chi-square test, for which the p-value was 0.13.

For Mr. Energy to know if the lollipop study was a good idea, an accurate understanding of p-values is crucial for conducting and interpreting research. In trying to prove that there is a difference in patient satisfaction scores when a lollipop is given, Dr. Bone needs to test the "null hypothesis." A null hypothesis, in this case, is that there is no difference in the mean satisfaction score between the lollipop and non-lollipop groups. After comparing the mean scores with a t-test, the p-value was less than 0.0001. **The p-value is the probability of finding this result if the null hypothesis is actually correct.** In this study, there is a less than 0.01% chance that there is actually no difference in mean patient satisfaction scores between the two groups.

Conventionally, a p-value less than 0.05 has been used to indicate statistical significance. While this is the most commonly applied threshold for statistical significance, it is important to remember that this cutoff is not "absolute." Its importance must be weighed against the research methodology being employed as well as the degree of type 1 error (rejecting the null hypothesis when it is actually true) that is acceptable. P-values are often misinterpreted in various ways. For example, it is sometimes thought of as the probability that the null hypothesis is true. As noted previously, however, the actual definition is the probability that the current statistical result would be found if the null hypothesis is actually true. Furthermore, a p-value is not absolute; it must be interpreted in the context of the study design (and its limitations), the statistical analysis (and appropriate use of tests), and practical considerations regarding the alternative hypothesis. Finally, the p-value does not, in and of itself, convey the etiology of a statistical finding.

As a result of this study, Mr. H.I. Energy approved a budget line increase for lollipops in the pediatric orthopaedic clinic.

Buoyed by her recent institutional quality improvement initiative, Dr. Bone notices that the incidence of intoeing seems to be decreasing, and she decides to study a "new treatment" for management of children with intoeing. She is unsure how many patients are likely to benefit from this novel treatment and decides to embark on a

clinical trial to test the efficacy of this new super-secret cure for intoeing compared to the gold standard of observation. The primary outcome of interest is decreased failure rate, and secondary outcomes include improved range of motion, quality of life as well as decreased pain as measured by clearly defined parameters. The trial ran for 2 years with no loss to follow up.

Figure 2. *Intoeing: The bane of pediatric orthopaedists*

At the end of the 2-year study period, her statistician gave her the following results: "Of the 50 patients enrolled, the failure rates in patients treated with the new treatment and standard observation were 1% and 6%, respectively. The relative risk was 0.17, the absolute risk reduction was 5%, and number needed to treat 20 (95% CI 15 - 27)." This seemed like a lot of information, and Dr. Bone wasn't sure how she should explain this to her next 20 patients who are concerned about their child's intoeing and their future Olympic aspirations.

A key to practicing evidence-based medicine is weighing potential benefits against potential harms or risks to patients. Useful statistical measures to assess risk include relative risk and absolute risk. Both are measures of incidence whereby the former is **the ratio of new cases of a particular outcome in the treatment to new cases in the control groups. Absolute risk reduction (ARR) is computed as the incidence in the control group - incidence in the treatment (intervention) group.**[1-3] Realistically, not everyone is expected to benefit from a treatment or intervention. As such, **the number needed to treat (NNT) is a statistical measure particularly used in clinical trials to communicate the effectiveness/efficacy of a treatment, procedure, or intervention**.[1-3] It represents the average number of patients who need to be treated to prevent one additional bad outcome (or receive a benefit) over a given period. Mathematically, it is the inverse of the absolute risk reduction (ARR) (i.e., 1/ARR).

Based on the above scenario, the NNT of 20 means that 20 patients with intoeing would need to be treated with the new procedure for one additional patient not to have failed outcome within a 2-year period. The 95% confidence interval above means that in this case, 95% of the time, the NNT will fall within the range of 15 to 27 patients. In other words, Dr. Bone would explain to her future patients that for every 20 patients treated with the super-secret intoeing treatment rather than the observation (gold standard), one patient would achieve improved outcomes at 2 years. This information is valuable in light of the morbidity of the super-secret treatment and the natural history of the dreaded intoeing. If the super-secret treatment was minimal (e.g., Tootsie Pop), then parents might consider it an option. If the super-secret treatment was intrusive (e.g., tibial osteotomy), the cure would likely be considered worse than the problem in light of the relatively benign natural history.

Over the past few months, Dr. Bone has been seeing an increased number of children and adolescents in her pediatric orthopaedic clinic with a novel condition known as "spidermanism." Amazingly, this condition gives affected patients abilities resembling those of the Marvel Comics superhero, Spider-Man. All patients with spidermanism have an increased proportion of type II muscle fibers on muscle biopsy. Other signs of spidermanism include increased physical strength and speed, hypermobile joints, and a heightened fight-or-flight response. Some, but not all, patients with spidermanism can also

produce silk-like webbing from the volar surface of both wrists during periods of stress. This ability to produce webbing is unique to the spidermanism condition.

Figure 3. *Superfit superheroes (spidermanism) can be suspected if they have increased levels of type II muscle fibers (sensitive testing) and confirmed by their ability to develop webbing from their wrists (specific testing).*

When seeing a patient with suspected spidermanism, Dr. Bone performs one of two tests to confirm the diagnosis:

A. Perform a muscle biopsy to look for an increased proportion of type II muscle fibers.
B. Perform an exercise stress test and look for the production of webbing from the patient's wrists.

Test A is a diagnostic test that is highly sensitive as all patients with spidermanism have increased type II muscle fibers, while Test B is a highly specific diagnostic test as no other disorders will produce a web from the wrists.

Sensitivity is defined as the percentage of people who have a condition and also test positive out of the total number of people who have the condition. In other words, **sensitivity is the probability of correctly testing positive when the condition is present.** If this probability is very high, then we can conclude that when the diagnostic test returns a negative result, the condition is most likely absent. A commonly used acronym that is useful for remembering this concept is SnNOut: a diagnostic test with high sensitivity will rule OUT a condition if the test results are negative.[4]

Here, Test A investigates the proportion of type II muscle fibers as seen on biopsy, with a positive test result indicating an increased proportion of type II fibers and a negative test result indicating the absence of increased type II fibers. We can conclude that this test is highly sensitive; if there is a negative test result (i.e., there is not an increased proportion of type II fibers), then we can be confident that the patient does not have spidermanism, due to the fact that all spidermanism patients must display increased type II muscle fibers.

Can we conclude that the patient has spidermanism if the patient does have an increased proportion of type II muscle fibers? Not necessarily! Many other people, such as trained athletes, can also have an increased proportion of type II fibers for reasons unrelated to spidermanism.[5] Thus, Test A lacks specificity for detecting spidermanism.

Now, let us consider Test B, which is a highly **specific** test. Specificity is defined as the percentage of people who do not have a condition and also test negative out of the total number of people who do not have the condition. In other words, **specificity is the probability of correctly testing negative when the condition is absent.** If this probability is very high, then we can conclude that when the diagnostic test returns a positive result, the condition is most likely present. A commonly used acronym that is useful for remembering this concept is SpPIn: a diagnostic test with high specificity will rule IN a condition if the test results are positive.[4]

For Test B, a positive test result indicates the ability to produce silk-like webbing from the wrists during an exercise stress test. A negative test result indicates an inability to produce webbing during the stress test. We can conclude that this test is highly specific. If there is a positive test result (i.e., the patient can produce webbing),

then we can be confident that the patient has spidermanism due to the fact that webbing production can only occur in spidermanism patients.

On the other hand, can we conclude that the patient does not have spidermanism if there is a negative test result (i.e., no webbing produced during the stress test)? No, it is known that some patients with spidermanism do not have the ability to produce webbing from the wrists when stressed. Thus, unlike Test A, Test B is not very sensitive. A negative test result does NOT rule out the condition in question.

In summary, Dr. Bone has utilized two established tests that are different in their diagnostic properties for spidermanism: Test A is highly sensitive but not as specific, while Test B is highly specific but not as sensitive.

Dr. Bone has four children who have been complaining bitterly of neck pain and achiness lately. In parallel, Dr. Bone has noticed that her wireless account is constantly buffering while she is on her department meetings while her children are in Zoom classes, which they simultaneously seem to be watching a movie on an iPad while frequently *"Face-Snapping their Tweeters"* on their smartphones. She hypotheses that the recent increase in neck pain and the slow internet may be related. As an academic orthopaedist she develops a new test for "Tech Neck," a condition causing neck pain from looking down at phones, tablets, or computers for extended periods of time. With this test, she hopes to stop her children from complaining of neck pain while increasing her wireless speed at home.

The prevalence of Tech Neck at her local school is known to be 30%. She administers this test to 100 students; 30 would be expected to have Tech Neck while 70 would not. Of the 30 with Tech Neck, 24 test positive; thus, this test has a sensitivity of 80%. Of the 70 that don't have Tech Neck, the test is negative in 63; thus, this test has a specificity of 90%. Based on these values, she makes the table at right.

Figure 4. *Could Dr. Bone's son have a sore neck as a result of forced remote studying while balancing his ultra-important social media platform?*

The **positive predictive value (PPV) is the probability that a subject with a positive test actually has the disease.** For Dr. Bone's test, of the 31 who tested positive, 24 actually have Tech Neck for PPV of 77%. You can calculate the PPV by dividing the number of true positives by the number of true positives + number of false positives. In this case, it would be 24/(24+7).

The **negative predictive value (NPV) is the probability that a subject with a negative test truly does not have the disease.** For Dr. Bone's test, of the 69 who tested negative, 63 actually did not have Tech Neck for NPV of 91.3%. You can calculate the NPV by dividing the number of true negatives by the number of true negatives + number of false negatives. In this case, it would be 63/(63+6).

	Does the patient have the Tech Neck?			
Test result		Yes	No	Total
	Positive	24	7	31
	Negative	6	63	69
	Total	30	70	100

Fewer false positives and false negatives are signs of a good test. In an ideal situation with zero false positives and negatives, you would have a PPV and NPV of 100%. PPV and NPV are directly affected by the prevalence of disease in the population. As prevalence increases, the PPV will also increase. Conversely, the NPV will decrease as prevalence increases.

Summary

Dr. Bone has led us through the definitions of common statistical terms using various case studies from her very diverse clinical practice and family life. A correct understanding of these important terms will help us interpret the literature and design studies of our own which may or may not involve superheroes with excessive femoral anteversion and who appreciate a good sucker while limiting screen time.

POSNA Evidence-Based Practice Committee

Matthew Schmitz, MD, San Antonio Military Medical Center, San Antonio, TX

Tracey Bastrom, MA, Rady Children's Hospital, San Diego, CA

Arvindera Ghag, MD, FRCSC, BC Children's Hospital, Vancouver, BC

Joseph Janicki, MD, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL

Judson Karlen, MD, Phoenix Children's Hospital, Phoenix, AZ

Susan Nelson, MD, MPH, University of Rochester Medical Center, Rochester, NY

Maegen J. Wallace, MD, Children's Hospital and Medical Center, Omaha, NE

Indranil Kushare, MD, Baylor College of Medicine & Texas Children's Hospital, Houston, TX

Ronald Lewis, MD, Pediatric Orthopedics of Charleston, Charleston, SC

R. Justin Mistovich, MD, Rainbow Babies and Children's Hospital, Cleveland, OH

William Phillips, MD, Baylor College of Medicine & Texas Children's Hospital, Houston, TX

Kelly Vanderhave, MD, Carolinas Medical Center, Charlotte, NC

Unni Narayanan, MD, FRCSC, The Hospital for Sick Children, Toronto, ON

Scott D. McKay, MD, Baylor College of Medicine & Texas Children's Hospital, Houston, TX

Acknowledgments

Images printed with permission from Texas Children's Hospital, Houston, TX.

References

1. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med. 1988;318(26):1728–33.

2. Citrome L, Ketter TA. When does a difference make a difference? Interpretation of number needed to treat, number needed to harm, and likelihood to be helped or harmed. Int J Clin Pract. 2013;67(5):407–11.

3. Cook D, Sackett D. The number needed to treat: a clinically useful measure of treatment effect. BMJ. 1995;310:452–4.

4. Pewsner, D. et al. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. BMJ 329, 209–213 (2004).

5. Wilson, J. M. et al. The Effects of Endurance, Strength, and Power Training on Muscle Fiber Type Shifting. J. Strength Cond. Res. 26, 1724–1729 (2012).