Beyond SpPIN and SnNOUT: Considerations with Dichotomous Tests During Assessment of Diagnostic Accuracy

Eric J Hegedus, PT, DPT, MHSc[1]; Ben Stern, PT, DPT, MS[2]

In an effort to *rule in* or *rule out* a disease or condition, clinicians often use one or multiple tests to confirm or refute a hypothesis. Specific tests are advocated as highly sensitive or specific, purportedly rendering that test effective at either *ruling out* (if sensitive) or *ruling in* (if specific) a given diagnosis. Unfortunately, using sensitivity or specificity independently to confirm a diagnosis may lead to error. The purpose of this clinimetrics corner titled "Beyond SPIN and SNOUT: Considerations with Dichotomous Tests During Assessment of Diagnostic Accuracy" is to discuss the limitations of these shortcuts for diagnostic accuracy.

Sensitivity, or the *true positive rate*, is defined as the proportion of patients who have the disease that the test identifies as positive and is calculated using the formula: Sn = True positives/(True positives + False negatives). Specificity, or the *true negative rate*, is defined as the proportion of patients who do not have the disease that the test identifies as negative. Specificity is calculated using the formula: Sp = True negatives/(True negatives + False positives). Both sensitivity and specificity are essential when describing the value of selected tests and measures used during diagnosis.

Sackett et al[1] introduced SpPIn (when specificity is high, a positive test rules in the diagnosis) and SnNOut (when sensitivity is high, a negative test rules out the diagnosis); and since this introduction, many practitioners have adopted these mnemonics as irrefutable standards. Despite the promotion by Sackett, the use of these descriptors has been questioned. In 2004, Pewsner et al[2] urged caution in the unqualified use of SpPIn and SnNOut, warning that potential weaknesses were present behind the singular use of SpPIn and SnNOut that might have unwanted clinical consequences.

ABSTRACT: Paramount to efficient and effective care is the determination of an accurate diagnosis that leads to the proper referral and/or intervention. In an effort to improve the clinical utility of diagnostic accuracy calculations, researchers have promoted the use of the mnemonics SpPIn (if specificity is high, a positive test *rules in* pathology) and SnNOut (if sensitivity is high, a negative test *rules out* pathology). Using examples from diagnostic accuracy studies and a review of pertinent literature, this clinimetrics corner outlines additional considerations for clinicians when consuming research in this area. The paper has three foci. First, sensitivity, specificity, and other estimates of the diagnostic accuracy of dichotomous physical examination tests should be viewed as estimates with confidence when those estimates are expressed as confidence intervals. Second, appropriate power must be considered when evaluating each study. Last, the quality of a diagnostic accuracy study can affect the generalizibility of the results to practice environments.

KEYWORDS: Bayes Theorem, Diagnostic Accuracy, Likelihood Ratio, QUADAS, Sensitivity, Specificity

SpPIn, SnNOut, and Confidence Intervals

The temptation exists to simplify the diagnostic process and view the estimates of sensitivity or specificity as single numbers when, in fact, each is actually an estimate of certainty[3,4]. In most diagnostic studies, values bound within a 95% confidence interval (CI) are considered acceptable measures. The wider the CI, the less precise is the presentation of the data. The width of the CI is determined by 1) sample size, 2) variability between and within the subjects being studied, and 3) the level of confidence desired. The reason that the CI is important when using SpPIn and SnNOut is that these mnemonics take continuous data such as sensitivity and specificity scores and dichotomize that data into a specific value, often termed as "high" or "low." Sensitivity or specificity values that qualify as "high" are usually defined as scores of 95% or greater, especially in a low prevalence (a

[1]Associate Professor and Vice-Chief, Doctor of Physical Therapy, Duke University Medical Center, Durham, NC
[2]Physical Therapist, 360 Physical Therapy, Fountain Hills, AZ
Address all correspondence and requests for reprints to : Dr. Eric J Hegedus, eric.hegedus@duke.edu

few individuals with the disease) setting[5]. However, an underpowered study or one that has a small sample size may identify a single sensitivity value as 95% (acceptable for ruling out a diagnosis when the test is negative) when the lower end of the confidence interval may dip into the 80's (not acceptable for ruling out a diagnosis when the test is negative). The wider the confidence interval, the less precise the single measure of accuracy.

In a recently published study[6], knee joint line tenderness was found to have a sensitivity of 95%, meeting the suggested value for *ruling out* a torn tibial meniscus when a clinical finding for tenderness is negative. However, the calculated CI was .87–.98, with the lower end of the CI falling below the recommended value for *ruling out* a condition. This substantiates weaknesses outlined previously[6] and exposes two significant areas of consideration of diagnostic accuracy calculations. First, the importance of the sample size (power) cannot be overstated, and second, that single measures of accuracy have significant limitations. Consequently, additional calculations such as likelihood ratios (LR) demonstrate greater clinical utility because the calculations incorporate both sensitivity and specificity.

Power

A hallmark of solid research design is an a priori estimate of the number of subjects needed to detect a significant relationship when one is present[7]. This a priori estimate is called a power estimate[7]. Using the previous example of detecting a torn tibial meniscus, clinical intuition may provide a suggestion that joint line tenderness is generally a sensitive test, but without appropriate power in a study design, one cannot confirm this suspicion. Increasing the power of the study and involving more subjects who are representative of the population in general can improve confirmation of a test's utility. Unfortunately, most diagnostic accuracy studies are woefully underpowered.

Power calculations for diagnostic accuracy studies were first introduced in 1991[3]. Recently, Flahaut et al[8] published tables that allow easier estimation of necessary sample sizes based on a required/expected sensitivity or specificity findings and a desired lower 95% confidence limit. Using the values outlined by Flahaut and colleagues, and the study data on joint line tenderness provided by Grifka et al[6] (which involved a sample size of 113), we can calculate the necessary sample size needed to meet a pre-desired lower confidence interval. If the desired sensitivity was 95% and the desired lower confidence limit was 90%, the actual sample size for Grifka et al[6] would have required 298 subjects. Sufficient power allows certainty that the sensitivity and specificity reported for a test will be the same as those plausible in a traditional clinical practice.

An important note is that sensitivity and specificity do have further limitations. First, and most important, is that sensitivity and specificity are parameters reported in a diagnostic accuracy study that are based on knowing whether or not the patient has the pathology. Clinically, this information is unavailable and presumably, the test is being performed either to detect or to rule out the presence of disease. Next, there is generally a trade-off between sensitivity and specificity in that as one parameter rises, the other falls. This co-dependent relationship makes the ranking of the usefulness of a special test less than intuitive and makes the use of either one without the other (as in SpPIn and SnNOut) potentially misleading. Further, a sensitivity or specificity estimate is dependent on the pre-test probability or prevalence of the pathology[9] and the severity of that pathology[10]. In other words, the estimates of sensitivity and specificity produced by a given study are affected by the number of diseased subjects in that study and the severity of the disease. An additional disadvantage of sensitivity is that it applies only to individuals who have the disease, while specificity is applicable only to those who do not[11]. Finally, sensitivity and specificity cannot easily be used to convert a pre-test probability of a disease to a post-test probability of disease[10]. Fortunately, likelihood ratios (LR), which combine sensitivity and specificity, alleviate many of the shortcomings of sensitivity and specificity and SpPIn and SnNOut.

Likelihood Ratios, Bayes Theorem, and Nomograms

A practical way of looking at LRs are as modifiers of probability. A practitioner performing a physical examination should be interested in whether each portion of the examination moves him or her closer to or further away from a diagnosis and the magnitude of that change[10]. A positive LR (LR+), represented by the formula sensitivity/(1-specificity), moves the practitioner closer to a diagnosis while a negative LR (LR-), represented by the formula, (1-sensitivity)/specificity, moves the practitioner further away from a diagnosis. The magnitude of the change is ranked on a 0 to infinity scale, with 1.0 indicating a test that provides no modification of post-test probability, and therefore is of questionable value. The direction and magnitude of change is captured via Bayes theorem and represented by the formula Pre-test Odds x LR = Post-test Odds.

The following example involving pre-test probability, LRs, and Bayes theorem may further elucidate the utility of these tools. Consider the case of an overweight patient who presents with a chief complaint of knee pain and slow onset of swelling that commenced after twisting on a planted lower extremity. A clinician may estimate the probability of this patient having a torn meniscus as 40%, based on reports in research, the practitioner's experience, or some combination of the two[10,12]. The clinician performs palpation of the knee and finds an absence of joint line tenderness. Is the absence of a positive finding in this case compelling enough to *rule out* the presence of a torn meniscus?

Using Bayes theorem, one can calculate the capacity of this assumption. Because Bayes theorem deals with odds and not probability, the probability values require conversion to an odds ratio. The 40% (.40) pre-test probability is converted to .667 odds [.40/ (1-.40)] and multiplied by the likelihood ratio, in this case, the LR-, which equals 1.0 (value used from the reported values of Grifka

et al[6]). The resulting calculation is a post-test odds of .667. The post-test odds are then converted back to post-test probability through the formula [.667/(1+.667]. Even without converting back to post-test probability, the reader has no doubt realized that the pre-test probability was completely unchanged by the negative joint line tenderness test with a LR- of 1.0. In this example, SnNOut does not rule out a torn meniscus and highlights the risk of strict application of the SpPIn and SnNOut mnemonics and the importance of using LRs that combine sensitivity and specificity. Also acknowledged in this example is the reason why LRs and Bayes theorem are not used regularly by practitioners[13,14]. Conversion from probability to odds and back again is cumbersome[14].

An alternative to odds ratio conversion was produced by Fagen in 1975[15]. Use of a nomogram allows the multiplicative analysis of pre-test probability and an LR to calculate post-test probability, without requiring a conversion to an odds ratio. Post-test probability calculations are associated with the probability of having or not having the condition when a test is positive (LR+) or negative (LR-). To improve the clinical utility further, McGee[14] developed a table linking LRs to approximate percentage changes in post-test probability (Table 1). McGee recommended that practitioners memorize 2, 5, and 10. With an LR+ of 2, 5, and 10, the percent increase in post-test probability is 15, 30, and 45, respectively. With an LR- , the practitioner uses the inverse version of the numbers or 1/2 (.50), 1/5 (.20), and 1/10 (.10) so that the percent decrease in post-test probability is 15, 30, and 45.

Quality and Bias

As was illustrated using the example from Grifka et al[6], SnNOut may be hampered by weaknesses when the specificity is low; thus, both sensitivity and specificity should be considered a concern when analyzing test results. Yet, what may be more intriguing are the findings of a recent review[16] of 14 articles that examined the joint line tenderness test. In this review the sensitivities, specificities, and LRs were grossly disparate among the 14 articles, with LR+ values that ranged from .86 to 36! How can joint line tenderness be so valuable in one study[17] that it could serve as a stand-alone test to diagnose a torn meniscus and yet, in another study[18], be so poorly diagnostic as to be worthless? The answer lies in the quality of design elements within each of the studies. Unfortunately, all diagnostic studies are not equal with regards to quality of design.

Higher-quality studies are those that reduce or minimize bias, defined as systematic error in the design or conduct of a research study[7]. In trials of intervention, many biases are addressed by randomizing patient group assignment. However, studies of diagnostic accuracy and the assessment of their quality are unique[19] and do not involve randomization of patients. Indeed, diagnostic accuracy studies require specific tools for measuring internal bias that are exclusive to this study design.

To date, most diagnostic accuracy study designs have demonstrated poor quality[20,21]. Prompted by this information, Whiting et al[19] produced the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) document, which is the first validated tool to aid a reader of research in critically assessing the quality of that research. The QUADAS tool consists of 14 questions that address the most common biases affecting internal validity (how the study was designed and implemented) and external validity (applicability of the study to daily practice) of diagnostic accuracy studies. These 14 questions were developed using Delphi mechanisms. A detailed discussion of all of the types of bias and their effect on estimates of sensitivity/specificity and LR+/LR- are beyond the scope of this paper but have been carefully outlined by Cook and colleagues[22].

In addition to the biases outlined in the QUADAS document, one additional area of bias is possible. A form of case-control design, frequently used by diagnostic accuracy studies, allows a potentially dramatic overestimation of the accuracy of a test[23,24]. When the control group without the condition of interest consists of individuals who are healthy or who have a condition non-related to the targeted pathology, results are often over-estimated. Calculations result in inflated, misleading values that would not occur in a situation of diagnostic uncertainty[7]. Consequently, a case-control design that retains diagnostic uncertainty (for example, comparing the finding of patients who have meniscus tears versus patellofemoral pain syndrome) is

TABLE 1. Likelihood ratios and bedside estimates

*Likelihood Ratio	Approximate Change in Probability (%)
Values between 0 and 1 decrease the probability of disease (LR-)	
.1	−45
.2	−30
.3	−25
.4	−20
.5	−15
1	0
Values greater than 1 increase the probability of disease (LR+)	
2	+15
4	+25
5	+30
6	+35
8	+40
10	+45

Adapted from McGee[14] with permission

essential for true assessment of test quality.

To illustrate the effect of bias, data from a recent systematic review[21] of an index test, the Active Compression Test, is an excellent choice for presentation. The Active Compression Test was originally developed with the intent of having a physical examination test that distinguished between shoulder pain that was either caused by an acromioclavicular pathology or from a superior labral anterior to posterior (SLAP) tear[25]. The original article[25] produced a great deal of excitement in the orthopedic community as a SLAP lesion is a difficult physical diagnosis. The study's sensitivity and specificity were reported as 99% and 98%, respectively. and LR+ and LR- were reported as 49.5 and .01, respectively. The Active Compression appeared to be a stand-alone test, useful in diagnosis of a SLAP lesion. Recently, higher-quality studies[26-30] estimated the likelihood ratios as close to 1, suggesting the index test made little or no modification of post-test probability. If one analyzes the quality of the studies using the QUADAS tool, the original publication demonstrates substantial internal biases that could have resulted in the higher reported values[31].

Final Considerations

This clinimetrics corner has treated test results as dichotomous, which means that there are two outcomes to the test, usually positive (producing pain, paresthesia, apprehension) or negative (no symptoms or change in symptoms). Dichotomous outcomes are the prevailing norm in orthopedic manual therapy. However, many practitioners realize that this is not often the case, especially if one relies on the concept of "concordant sign." Simply put, eliciting the concordant sign means that the chosen test or tests reproduce the symptom or symptoms that the patient has identified as the reason he or she sought a physical therapy consultation[32]. Frequently, the patient responds in the affirmative or the negative but we have all experienced the equivocal response of "not quite sure" or "maybe"[33]. Despite this clinical probability, in two systematic reviews of shoulder and knee physical examination special tests, intermediate or indeterminate results were rarely reported[16,21]. We believe that the reason for this rare report is that equivocal results are statistically inconvenient, and researchers subjectively dichotomize those responses to fit neatly into a 2x2 table (the basic table from which all estimates of diagnostic accuracy are calculated). The effect of this bias on estimates of diagnostic accuracy is not well known.

One final clinical consideration is that physical exam tests are rarely used in isolation. Normally, a physical therapist would start with history and systems review and then move through several steps of a physical examination including some combination of motion assessment, strength, palpation, and special tests. Each step of the examination process has its own associated likelihood ratio or ratios, and when the examination is sequential, the post-test probability of one test becomes the pre-test probability of the next test. The earlier example of an overweight patient who presented with chief complaints of knee pain and slow onset of swelling after twisting on a planted lower extremity had a mythical pre-test probability of 40% of having a torn meniscus. Recall that 40% pre-test odds were equal to .667. The hypothetical clinician then performs motion testing and notices that the patient's painful knee has limited flexion with pain produced at end range, a test with an LR+ of 1.6[34]. The clinician then chooses the Dynamic test[35] with an LR+ of 8.5; a positive test would improve the post-test probability as follows: .667 x 1.6 = 1.067 (51% post-test probability), then 1.067 x 8.5 = 9.0695 (90% post-test probability.

This example demonstrates how a sequential examination builds a preponderance of evidence to support a diagnosis. With understanding of the rest of this article, the reader should question the LRs presented here since no CIs or reliability estimates were presented, the studies from which the LRs were taken were underpowered and have some design faults that produce bias, and the case description provides little insight as to whether detecting a torn meniscus will change the management of the patient. These are the key elements that help the clinician realize that estimates of diagnostic accuracy parameters are just that—an estimate. This realization will help take the clinician far beyond SpPIn and SnNOut.

REFERENCES

1. Sackett D, Haynes R, Guyatt G, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Boston, MA: Little Brown, 1992.
2. Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": A note of caution. *BMJ* 2004;329:209–213.
3. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: Sample size estimation for diagnostic test studies. *J Clin Epidemiol* 1991;44:763–770.
4. Harper R, Reeves B. Reporting of precision of estimates for diagnostic accuracy: A review. *BMJ* 1999;318:1322–1323.
5. Obuchowski NA, Graham RJ, Baker ME, Powell KA. Ten criteria for effective screening: Their application to multislice CT screening for pulmonary and colorectal cancers. *AJR Am J Roentgenol* 2001;176:1357–1362.
6. Grifka J, Richter J, Gumtau M. Clinical and sonographic meniscus diagnosis. *Orthopade* 1994;23:102–111.
7. Kocher MS, Zurakowski D. Clinical epidemiology and biostatistics: A primer for orthopaedic surgeons. *J Bone Joint Surg Am* 2004;86–A:607–620.
8. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005;58:859–862.
9. Fritz JM, Wainner RS. Examining diagnostic tests: An evidence–based perspective. *Phys Ther* 2001;81:1546–1564.
10. Bhandari M, Guyatt GH. How to appraise a diagnostic test. *World J Surg* 2005;29:561–566.
11. Tatsioni A, Zarin DA, Aronson N, et al. Challenges in systematic reviews of diagnostic technologies. *Ann Intern Med* 2005; 142:1048–1055.

12. Elstein AS, Schwartz A. Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *BMJ* 2002;324:729–732.
13. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet* 2005;365:1500–1505.
14. McGee S. Simplifying likelihood ratios. *J Gen Intern Med* 2002;17:646–649.
15. Fagan TJ. Letter: Nomogram for Bayes theorem. *N Engl J Med* 1975;293:257.
16. Hegedus EJ, Cook C, Hasselblad V, Goode A, McCrory DC. Physical examination tests for assessing a torn meniscus in the knee: A systematic review with meta-analysis. *J Orthop Sports Phys Ther* 2007;37:541–550.
17. Eren OT. The accuracy of joint line tenderness by physical examination in the diagnosis of meniscal tears. *Arthroscopy* 2003; 19:850–854.
18. Noble J, Erat K. In defence of the meniscus: A prospective study of 200 meniscectomy patients. *J Bone Joint Surg Br* 1980;62–B: 7–11.
19. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess* 2004;8:25.
20. Sheps SB, Schechter MT. The assessment of diagnostic tests: A survey of current medical research. *JAMA* 1984;252:2418–2422.
21. Hegedus EJ, Goode A, Campbell S, et al. Physical examination tests of the shoulder: A systematic review with meta-analysis of individual tests. *Br J Sports Med* 2007:42:80–92.
22. Cook C, Cleland J, Huijbregts P. Creation and critique of studies of diagnostic accuracy: Use of the STARD and QUADAS methodological quality assessment tools. *J Man Manipulative Ther* 2007;15:93–102.
23. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999; 282:1061–1066.
24. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469–476.
25. O'Brien SJ, Pagnani MJ, Fealy S, McGlynn SR, Wilson JB. The active compression test: A new and effective test for diagnosing labral tears and acromioclavicular joint abnormality. *Am J Sports Med* 1998;26:610–613.
26. Parentis MA, Glousman RE, Mohr KS, Yocum LA. An evaluation of the provocative tests for superior labral anterior posterior lesions. *Am J Sports Med* 2006;34:265–268.
27. Myers TH, Zemanovic JR, Andrews JR. The resisted supination external rotation test: A new test for the diagnosis of superior labral anterior posterior lesions. *Am J Sports Med* 2005;33:1315–1320.
28. Nakagawa S, Yoneda M, Hayashida K, Obata M, Fukushima S, Miyazaki Y. Forced shoulder abduction and elbow flexion test: A new simple clinical test to detect superior labral injury in the throwing shoulder. *Arthroscopy* 2005;21:1290–1295.
29. McFarland EG, Kim TK, Savino RM. Clinical assessment of three common tests for superior labral anterior-posterior lesions. *Am J Sports Med* 2002;30:810–815.
30. Morgan CD, Burkhart SS, Palmeri M, Gillespie M. Type II SLAP lesions: Three subtypes and their relationships to superior instability and rotator cuff tears. *Arthroscopy* 1998;14:553–565.
31. Cook C, Hegedus E. *Orthopaedic Clinical Examination Tests: An Evidence-Based Approach.* Upper Saddle River, NJ: Prentice Hall, 2008.
32. Cook C. *Orthopedic Manual Therapy: An Evidence-Based Approach*. Upper Saddle River, NJ: Prentice Hall, 2007.
33. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Med Decis Making* 1987;7:107–114.
34. Fowler PJ, Lubliner JA. The predictive value of five clinical signs in the evaluation of meniscal pathology. *Arthroscopy* 1989; 5:184–186.
35. Mariani PP, Adriani E, Maresca G, Mazzola CG. A prospective evaluation of a test for lateral meniscus tears. *Knee Surg Sports Traumatol Arthrosc* 1996;4:22–26.