Information in practice

Ruling a diagnosis in or out with “SpPIn” and “SnNOut”: a note of caution

Daniel Pewsner, Markus Battaglia, Christoph Minder, Arthur Marx, Heiner C Bucher, Matthias Egger

Dr X is back from her annual leave. Dr Y, the locum doctor, reports on the patients she saw during her absence, including a 40 year old teacher who had sprained her right ankle. Returning from a conference, she had stumbled while walking down the stairs with a heavy bag. Examination revealed a moderately swollen lateral right ankle. The patient was able to walk but was clearly in pain. Her breath smelt of alcohol.

Ruling diagnoses in and out with SpPIns and SnNOuts

Dr Y had applied the Ottawa ankle rules—decision rules designed to exclude fractures of the malleolus and the midfoot—and found no bone tenderness.1 He had previously visited the website of a centre for evidence based medicine2 and printed out a list of diagnostic tests that can rule out, or rule in, the condition in question without requiring further investigations.

The probability of disease, given a positive or negative test result (post-test probability), is usually obtained by calculating the likelihood ratio of the test result and using formulas based on Bayes’s theorem (see box 1), or a nomogram,3 to convert the estimated probability of the suspected diagnosis before the test result was known (pretest probability) into a post-test probability, which takes the result into account.4 Likelihood ratios indicate how many times more likely a test result is to be expected in a patient with the disease compared with a person free of the disease and thus measure a test’s ability to modify pretest probabilities.

David Sackett and others have argued that such calculations are unnecessary when a test is highly sensitive or highly specific.5,6 In this situation the likelihood ratio of a negative test will generally be very small, and the likelihood ratio of a positive test very large. A negative test will thus rule out, and a positive result rule in, disease. Two mnemonics that capture the properties of such tests have been coined: SnNOut (high sensitivity, negative, rules out) and SpPIn (high specificity, positive, rules in).7 This concept has become increasingly popular, with many websites for evidence based medicine listing such tests and inviting users to nominate further SpPIn and SnNOut tests. The understanding of the SnNOut principle among medical students was recently examined in a randomised trial.7

<table>
<thead>
<tr>
<th>Summary points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative results from highly sensitive tests can rule a diagnosis out (negative, severe) and positive results from highly specific tests can rule a diagnosis in (specific, positive, in = SpPIn)</td>
</tr>
</tbody>
</table>

Studies quoted as showing SpPIn or SnNOut properties may be affected by spectrum bias, partial verification bias, or incorporation bias. Others may be too small to define test characteristics with sufficient precision

The power of a test to rule a diagnosis out does not depend exclusively on its sensitivity, as suggested by the SnNOut rule, but is reduced by low specificity. Similarly, the power to rule in depends on both specificity and sensitivity

The evidence from studies of a test’s accuracy should be critically assessed, and post-test probabilities (with 95% confidence intervals) should be calculated when evaluating potential SnNOut or SpPIn tests

Assuming that a diagnosis can be ruled in or out with confidence, when in reality it cannot, could have serious consequences for patients

The website listed the Ottawa ankle rules as a SnNOut test,8 indicating that in the teacher’s case a fracture could safely be ruled out without radiography. Indeed, the patient made an uneventful and full recovery within four weeks. Alarmed by the patient’s alcoholic breath, Dr Y wondered whether an alcohol problem might have contributed to the accident and used the CAGE questions (see box 2) to investigate this further. According to the same website,8 the CAGE instrument has SpPIn properties, ruling the diagnosis in if two or more questions are answered affirmatively. The patient confirmed that she felt she should cut down on alcohol and that she had felt had repeatedly about her drinking. Dr X, who had known her for over 10 years, explained that the patient’s alcohol intake was moder-
### Box 1: Definitions of concepts and terms

- **Sensitivity**—The proportion of people with the disease who are correctly identified by a positive test result (true positive rate).
- **Specificity**—The proportion of people free of the disease who are correctly identified by a negative test result (true negative rate).
- **SnNOut**—Mnemonic to indicate that a negative test result (N) of a highly sensitive test (Sn) rules out the diagnosis (Out).
- **SpPIn**—Mnemonic to indicate that a positive test result (P) of a highly specific test (Sp) rules in the diagnosis (In).
- **Likelihood ratio**—Measure of a test result's ability to modify pretest probabilities. Likelihood ratios indicate how many times more likely a test result is in a patient with the disease compared with a person free of the disease.
- **Likelihood ratio of a positive test result (LR+)**—The ratio of the true positive rate to the false positive rate: sensitivity/(1 − specificity).
- **Likelihood ratio of a negative test result (LR−)**—The ratio of the false negative to the true negative rate: (1 − sensitivity)/specificity.
- **Pretest probability (prevalence)**—The probability that an individual has the target disorder before the test is carried out.
- **Post-test probability**—The probability that an individual with a specific test result has the target condition (post-test odds/[1 + post-test odds]) or (pretest probability/[1 + pretest probability]) × LR.
- **Pretest odds**—The odds that an individual has the target disease before being tested (pretest odds × LR).
- **Positive predictive value (PPV)**—The proportion of individuals with positive test results who have the target condition. This equals the post-test probability given a positive test result.
- **Negative predictive value (NPV)**—The proportion of individuals with negative test results who do not have the target condition. This equals one minus the post-test probability given a negative test result.

### Critical appraisal of test evaluation studies

In this article, we examine examples of test evaluation studies that websites and a textbook of evidence based medicine have cited as showing that the tests had SpPIn or SnNOut properties. The studies were chosen to illustrate methodological issues. We assessed the quality of studies as described elsewhere, extracted the two-by-two table from the original publication, and calculated likelihood ratios and post-test probabilities for various sensitivity and specificity levels.

### Box 2: CAGE questionnaire for detecting alcohol misuse

1. Have you ever felt you should Cut down on your drinking?
2. Have people Annoyed you by criticising your drinking?
3. Have you ever felt bad or Guilty about your drinking?
4. Have you ever had a drink first thing in the morning to steady your nerves or to get rid of a hangover (Eye opener)?
Table 1 Characteristics of test evaluation studies interpreted as demonstrating SnNOut properties (high sensitivity and negative result rules out)

<table>
<thead>
<tr>
<th>Test (diagnosis)</th>
<th>Reference test</th>
<th>Setting and patients</th>
<th>Results (tp/fp/tn/fn)</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Likelihood ratio (95% CI)</th>
<th>Pretest probability (95% CI)</th>
<th>Post-test probability (95% CI)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ottawa ankle rules (ankle or midfoot fractures)</td>
<td>Radiography</td>
<td>Consecutive adults in emergency departments</td>
<td>70/210/175/35</td>
<td>99% (92% to 100%)</td>
<td>38% (34% to 45%)</td>
<td>0.04 (0.01 to 0.3)</td>
<td>17.1%</td>
<td>0.7% (0.02% to 4.9%)</td>
<td>Post-test probability of 3% seems acceptable*</td>
</tr>
<tr>
<td>History of ankle swelling (ascites)</td>
<td>Ultrasound scan</td>
<td>Men admitted to a general medicine ward</td>
<td>14/16/170/3</td>
<td>93% (68% to 100%)</td>
<td>67% (52% to 80%)</td>
<td>0.10 (0.01 to 0.7)</td>
<td>23.8%</td>
<td>3.0% (0.1% to 15.8%)</td>
<td>Study compatible with post-test probability of 15.8%, but even the point estimate of 3% may be considered high in this situation</td>
</tr>
<tr>
<td>Loss of spontaneous retinal vein pulsation (increased intracranial pressure)</td>
<td>Clinical evidence, including lumbar pressure in 8 patients</td>
<td>Patients known to have increased pressure compared with individuals without suspected increased pressure</td>
<td>43/18/0/128</td>
<td>100% (92% to 100%)</td>
<td>88% (81% to 93%)</td>
<td>0.0 (0.0 to 0.2)</td>
<td>22.8%</td>
<td>0 (0 to 2.3%)</td>
<td>Diagnostic case-control study, vulnerable to spectrum bias. Results not reliable</td>
</tr>
<tr>
<td>NINCDS-ADRDA criteria (Alzheimer's disease)</td>
<td>Histology</td>
<td>Patients in secondary care</td>
<td>28/20/0/6</td>
<td>93% (78% to 99%)</td>
<td>23% (9% to 44%)</td>
<td>0.3 (0.06 to 1.3)</td>
<td>53.6%</td>
<td>25% (3.2% to 65.1%)</td>
<td>Power to rule diagnosis out based on low specificity, resulting in high post-test probability</td>
</tr>
<tr>
<td>Urinary albumin: creatinine ratio &gt;1.8 g/mol (microalbuminuria in diabetes)</td>
<td>Timed overnight albumin excretion</td>
<td>Men with diabetes screened for microalbuminuria in primary care</td>
<td>405/51/24/6</td>
<td>94% (92% to 98%)</td>
<td>93% (91% to 95%)</td>
<td>0.06 (0.04 to 0.1)</td>
<td>36.6%</td>
<td>3.4% (2.2% to 5.0%)</td>
<td>Post-test probability of ≤5% is acceptable in this situation</td>
</tr>
</tbody>
</table>

*tp=true positive, fp=false positive, fn=false negative, tn=true negative.
*Confirmed in recent meta-analysis of 27 studies.16
†Clinical criteria by National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and Alzheimer’s Disease and Related Disorders Association (ADRA).15

Table 2 Characteristics of test evaluation studies interpreted as demonstrating SpPIn properties (high specificity and positive result rules in)

<table>
<thead>
<tr>
<th>Test (diagnosis)</th>
<th>Reference test</th>
<th>Setting and patients</th>
<th>Results (tp/fp/tn/fn)</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Likelihood ratio (95% CI)</th>
<th>Pretest probability (95% CI)</th>
<th>Post-test probability (95% CI)</th>
<th>Comment: can positive test rule diagnosis in?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auscultatory percussion (pleural effusion)</td>
<td>Radiography</td>
<td>Patients with effusion compared to patients from surgical and medical wards</td>
<td>113/3/5/175</td>
<td>96% (90% to 99%)</td>
<td>100% (98% to 100%)</td>
<td>— (21 to —)</td>
<td>40%</td>
<td>100% (97.4% to 100%)</td>
<td>Diagnostic case-control study, vulnerable to spectrum bias. Results not reliable</td>
</tr>
<tr>
<td>23 positive answers in CAGE questionnaire*</td>
<td>Biomedical tests, medical record review, and physician interviews</td>
<td>Consecutive orthopaedic-medical patients. Reference test applied to sample of CAGE-negative patients only</td>
<td>60/15/7/400</td>
<td>51% (42% to 61%)</td>
<td>99% (98% to 100%)</td>
<td>206 (29 to 1468)</td>
<td>23%</td>
<td>98.4% (91.2% to 99.9%)</td>
<td>Results unreliable because of possible partial verification bias and wide confidence intervals</td>
</tr>
<tr>
<td>Abdominoojugular reflux (congestive heart failure)†</td>
<td>Clinical criteria that included reflux</td>
<td>Patients complaining of shortness of breath in emergency room</td>
<td>42/2/8/24</td>
<td>33% (10% to 65%)</td>
<td>94% (81% to 99%)</td>
<td>6.0 (1.3 to 29)</td>
<td>25%</td>
<td>66.1% (52.3% to 85.7%)</td>
<td>Results unreliable because of possible incorporation bias and wide confidence intervals. Low post-test probability</td>
</tr>
<tr>
<td>Third heart sound (heart failure)†</td>
<td>Echocardiography</td>
<td>Consecutive patients with suspected heart failure attending general practices</td>
<td>10/3/31/215</td>
<td>24% (12% to 40%)</td>
<td>99% (96% to 100%)</td>
<td>18 (5.1 to 63)</td>
<td>16%</td>
<td>76.9% (46.2% to 95.0%)</td>
<td>Power to rule diagnosis in is eroded by low sensitivity, resulting in low post-test probability. Wide confidence intervals</td>
</tr>
<tr>
<td>22 positive answers in CAGE questionnaire*</td>
<td>IC-10 criteria</td>
<td>Sample of black women admitted to trauma centre</td>
<td>25/12/28/57</td>
<td>83% (65% to 94%)</td>
<td>96% (93% to 98%)</td>
<td>21 (12 to 37)</td>
<td>9%</td>
<td>67.6% (50.2% to 82.0%)</td>
<td>Post-test probability too low, considering nature of diagnosis. Applicability to other settings and populations questionable</td>
</tr>
</tbody>
</table>

*tp=true positive, fp=false positive, fn=false negative, tn=true negative.
*See box 2 for CAGE criteria.
†Clinical criteria for the diagnosis of Alzheimer’s disease had SnNOut properties* based on a sensitivity of 93% (table 1).14 However, despite this high sensitivity, the likelihood ratio of a negative test was a modest 0.3, because of the test’s low specificity of 23% (100−93/23 = 0.3, see box 1). Indeed, in the population studied, the probability of Alzheimer’s disease, given a negative test, was 25% (table 1). The power to rule out a diagnosis thus depends on both sensitivity and specificity.

Similarly, the ability to rule in depends not only on specificity, as suggested by the SpPIn rule, but also on...
sensitivity. A study examining the presence of a third heart sound in the diagnosis of congestive heart failure (table 2)—which a website interpreted as demonstrating SpPIn properties—is an example of a highly specific test (99%) that suffers from a low sensitivity (24%). The figure shows how the power to rule a disease in or out is eroded when highly specific tests are not sufficiently sensitive, or highly sensitive tests are not sufficiently specific.

Transferability and applicability

The performance of a diagnostic test often varies considerably from one setting to another, which may be due to differences in the definition of the disease, the exact nature of the test, and its calibration and the characteristics of those with and without the disease in a given setting. For example, patients attending primary care practices will generally have disease at an earlier stage than patients in secondary and tertiary care, which may reduce a test's sensitivity. Patients free of the disease in tertiary care will tend to have other conditions, which could reduce the specificity of a diagnostic test. Interpreting data on a test's accuracy thus requires defining the exact nature of the test used, the disease, and the patient population studied. For example, the website that listed the CAGE questionnaire as a SpPIn test for alcohol dependence cited, as the evidence for this study, a study that had been performed in black women admitted to a trauma centre in the United States, which may not be applicable to other populations and settings.

Even when we assume that sensitivity and specificity do not change between settings and patient populations, test results will have different interpretations depending on whether a test is performed in a low risk population, such as in primary care, or high risk patients in a referral centre. For example, in the study evaluating the third heart sound in the diagnosis of heart failure, the pretest probability or prevalence in a general practice setting was 16%. In this situation, a positive test with a likelihood ratio of 18 will not allow the diagnosis to be ruled in with confidence: the post-test probability is only increased to 77% (table 2). If the pretest probability were 50%, however—in such as in a cardiology outpatient clinic—the same positive test would produce a post-test probability of 95% (see box 1 for formula).

The interpretation of studies will be strongly influenced by the nature of the condition and the invasiveness of further investigations. For example, a study assessing urinary albumin:creatinine ratios below 1.8 g/mol for ruling out microalbuminuria in men with type 2 diabetes in primary care and the study examining the absence of history of ankle swelling for ruling out ascites in men admitted to general internal medicine wards both produced post-test probabilities of about 3%. In the first case, we accepted a website's conclusion that the urinary albumin:creatinine ratio had SnNOut properties: we thought that the post-test probability of microalbuminuria was sufficiently low with a negative test result, considering that guidelines recommend regular testing of patients with type 2 diabetes. In the second case, however—and unlike the textbook—we thought that in men with suspected ascites but no history of ankle swelling a probability of ascites of 3%, a sign often associated with serious conditions, was still too high and that sonography should be used to rule the diagnosis in or out. As mentioned above, another problem with this study is the small sample size, which resulted in wide confidence intervals.

Conclusions

Prompted by a colleague's experience with the Ottawa ankle rule and the CAGE questionnaire, we examined diagnostic test evaluation studies, which on websites and in a textbook of evidence based medicine were interpreted as demonstrating tests' ability to conclusively rule a diagnosis in or out (SpPIn or SnNOut tests). We calculated likelihood ratios to measure the tests' power to rule the target conditions in or out and assessed the study designs for possible bias, and we have given examples of tests where these methodological issues raise questions about whether the tests have SpPIn or SnNOut properties.

We believe that the concept of SpPIn or SnNOut tests can help clinicians in interpreting diagnostic test results. However, identifying and promoting tests with SpPIn or SnNOut properties should be based on a careful appraisal of the evidence, including the methodological quality of the test evaluation studies, and not simply on the test's sensitivity or specificity. Likelihood ratios and typical post-test probabilities should be calculated and reported, together with measures of statistical uncertainty, as recommended in the Standards for Reporting Diagnostic Accuracy Studies (STARD). Assessments should ideally be based on a systematic review of all available studies, which may include a meta-analysis to increase the
precision of estimates of test accuracy. For example, a recent meta-analysis of 27 test accuracy studies of the Ottawa ankle rules confirmed that in many settings this decision aid can indeed exclude fractures and reduce the number of unnecessary radiographs. The result that results may not be transferable to other populations and settings should be stressed, and the information required to judge transferability and applicability should be provided. Clearly, assuming that a diagnosis can be ruled in or ruled out with confidence, when in reality it cannot, could have serious consequences for patients.

We thank Nicola Low for helpful comments on an earlier version of this manuscript.

Contributors: DP had the idea of critically appraising test accuracy studies that were interpreted as demonstrating SpPIn or SnNOut properties. CM advised on statistical issues. ME and DP wrote the first draft of the article. All authors contributed to the appraisal of the evidence from studies and to writing the final draft of the article.

Funding: The Krankenförsorgstiftung der Gesellschaft für das Gute und Gemeinnützige (GGG), Basel, Switzerland, and the Swiss Academy of Medical Sciences supported this study.

Competing interests: None declared.


2 Centre for Evidence-Based Medicine, University Health Network, Mount Sinai Hospital Toronto. Sensitivity & specificity (SnNouts and SpPins). www.cebm.utoronto.ca/glossary/spsn.htm (accessed March 2004).


15 Davies AP, Francis CM, Carter SA, Sutherland AR. Accuracy studies that were interpreted as demonstrating SpPIn or SnNOut. Arch Neurol 1978;35:37-40.


19 Centre for Evidence Based Medicine, University of Oxford. SpPIn and SnNOuts. www.cebm.ox.ac.uk/cebm/glossary/spsn.htm (accessed January 2003).


A sequence of errors

The weekly 500 mile aeroplane journey to the South African hospital shortened the working day for the visiting radiologist, so a car was always ready to take him straight from the airstrip to the waiting patients. If an examination under general anaesthesia was required, an anaesthetist would start to prepare the patient as soon as the aeroplane’s wheels touched the tarmac, so that the radiologist wasted no time at the hospital. The senior registrars from Cape Town, who took it in turns to provide the service, enjoyed the visit; it was good for radiological experience and even more for the change of scenery. Unfortunately I didn’t learn of the effect of the diagnosis on the patient’s outcome, but when, on the return journey to the airstrip, the chauffeur again incorrectly addressed me as “Professor” as I climbed into the large limousine. When I arrived at the hospital a patient had already anaesthetised, waiting for me to carry out a carotid angiogram to help diagnose a mass in the neck. I quickly washed up the vertebral artery on the injected side and down the one on the other side.

The vertebral arteries run up on each side of the neck deep to the carotids, passing from their origin in the subclavian arteries to the head, where they join to form the basilar artery. Erroroneously puncturing a vertebral artery at carotid angiography was unusual. Occluding the vertebral artery on the opposite side of the neck was totally unexpected. It meant that contrast material had passed up the vertebral artery on the injected side and down the one on the other side.

This retrograde flow indicated that there was an occlusion or stenosis at the origin of the contra lateral subclavian artery, and I was therefore able to diagnose instantly a subclavian steal syndrome on the opposite side of the neck to the injection. Closer clinical examination of the neck revealed that the mass on the ipsilateral side of the neck was due to obvious tuberculous adenopathy, and a carotid angiogram had not been indicated in the first place. In the twinkling of an eye, I had attempted an inappropriate examination, performed an unintended procedure, and diagnosed an unsuspected condition.

Unfortunately I didn’t learn of the effect of the diagnosis on the patient’s outcome, but when, on the return journey to the airstrip, the chauffeur again incorrectly called me professor the sequence of errors seemed complete.

Brian Witcombe consultant radiologist, Gloucestershire Royal Hospital, Gloucester