Academic Emergency Medicine
Official Journal of the Society for Academic Emergency Medicine

RESEARCH METHODS & STATISTICS

Understanding the Direction of Bias in Studies of Diagnostic Test Accuracy

Michael A. Kohn, MD, MPP, Christopher R. Carpenter, MD, MSc, and Thomas B. Newman, MD, MPH

Abstract

Ordering and interpreting diagnostic tests is a critical part of emergency medicine (EM). In evaluating a study of diagnostic test accuracy, emergency physicians (EPs) need to recognize whether the study uses case–control or cross-sectional sampling and account for common biases. The authors group biases in studies of test accuracy into five categories: incorporation bias, partial verification bias, differential verification bias, imperfect gold standard bias, and spectrum bias. Other named biases are either equivalent to these biases or subtypes within these broader categories. The authors go beyond identifying a bias and predict the direction of its effect on sensitivity and specificity, providing numerical examples from published test accuracy studies. Understanding the direction of a bias may permit useful inferences from even a flawed study of test accuracy.

ACADEMIC EMERGENCY MEDICINE 2013; 20:1194–1206 © 2013 by the Society for Academic Emergency Medicine

Diagnostic Tests (Versus Screening and Prognostic Tests)

Performing, ordering, and interpreting diagnostic tests plays a critical role in the practice of emergency medicine (EM).[1,2] Emergency physicians (EPs) are confronted with an increasing number of published studies evaluating the accuracy of new and old diagnostic tests. Approximately 63% of emergency department (ED) visits include laboratory tests, electrocardiograms (ECGs), or imaging studies,[3] but we also include in the term "diagnostic test" elements of the history or physical examination findings used by EPs to help determine patients' underlying conditions and guide clinical decisions about treatment or additional testing. Diagnostic tests are performed on patients with symptoms, as opposed to screening tests, which are performed on patients with no known symptoms.[4] Because symptoms are what bring patients into the ED, screening tests are generally of less interest in EM. Prognostic tests help predict incident outcomes rather than diagnose prevalent disease. These outcomes require a time interval to elapse and depend on additional random events that occur to the patient after the test is performed. We focus here on assessing the accuracy of a single diagnostic test for prevalent disease in symptomatic patients. While we limit this discussion to studies of test accuracy, we should point out that accuracy is not enough. An accurate test is primarily useful if correctly diagnosing the cause of a patient's illness improves outcomes by prompting effective treatment that would not have occurred otherwise.[5]

Studies of Diagnostic Test Accuracy

The prototypical study of diagnostic test accuracy compares a single index test to a gold standard determination of disease status. Disease status is assumed to be dichotomous: disease present (D+) or disease absent (D–). The index test can be dichotomous, multilevel, or continuous (Table 1). We will not be discussing test results that fall into more than two unordered categories.

For the most part, this discussion will assume that the index test is dichotomous (either positive or negative). For dichotomous tests, the most widely used indices of test accuracy are sensitivity and specificity. Sensitivity is the probability that a patient with the disease (a D+ patient) will have a positive test. Specificity is

From the Department of Epidemiology and Biostatistics, University of California at San Francisco (MAK, TBN), San Francisco, CA; the Emergency Department, Mills-Peninsula Medical Center (MAK), Burlingame, CA; and the Division of Emergency Medicine; Barnes Jewish Hospital, Washington University (CRC), St. Louis, MO.

Received April 25, 2013; revision received June 9, 2013; accepted June 16, 2013.

Dr. Carpenter, an associate editor for this journal, had no role in the peer review or publication decision for this manuscript. The authors have no disclosures or conflicts of interest to report.

Supervising Editor: Richard Sinert, DO.

Address for correspondence and reprints: Michael A. Kohn, MD, MPP; e-mail: michael.kohn@ucsf.edu.

ISSN 1069-6563
PII ISSN 1069-6563583

doi: 10.1111/acem.12255

Table 1
Types of Test Results

Type Result	Index Test	Target Condition	Gold Standard
Dichotomous	Elbow extension test (Unable to fully extend elbow? Yes/No)	Elbow fracture	Elbow radiograph
Dichotomous	PCR test for *B. pertussis* (positive/ negative)	Pertussis (whooping cough)	Pertussis culture
Multilevel (ordinal)	V/Q scan (normal, near normal, low probability, intermediate probability, high probability)	Pulmonary embolism	Pulmonary angiogram
Continuous	BNP (ranges from 0 to >1,300 pg/mL)	Acute heart failure	Consensus of two reviewing cardiologists

BNP = B-type natriuretic peptide; PCR = polymerase chain reaction; V/Q = ventilation/perfusion.

the probability that a patient without the disease (a D– patient) will have a negative test. Studies of test accuracy should also report likelihood ratios (LRs).[6,7] The general definition of the LR for a test result is the probability of obtaining that result in a patient *with* the disease divided by the probability of obtaining that result in a patient *without* the disease. The result need not be from a dichotomous test; it can be one of the several possible results from a multilevel test or a result in a particular interval from a continuous test.[8] For dichotomous tests, the LR of a positive result, LR(+), is sensitivity/(1 – specificity), and the LR of a negative result, LR(–), is (1 – sensitivity)/specificity. Other quantities commonly reported in studies of test accuracy are prevalence (the proportion of the population that has the disease) or, better yet, pretest probability (the prevalence in a clinical population of patients like the one currently being evaluated), positive predictive value (PPV, the probability that a person with a positive test has the disease), and negative predictive value (NPV, the probability that a person with a negative result does *not* have the disease). We elaborate on these quantities below in our discussion of sampling schemes and in Figure 1. We will not discuss the use of the term "accuracy" as the prevalence-weighted average of sensitivity and specificity, because we do not find this to be a useful quantity. If a disease is very rare with a prevalence of, say, 0.1%, then a test that is always negative has an accuracy of 99.9%

Bias in Studies of Test Accuracy

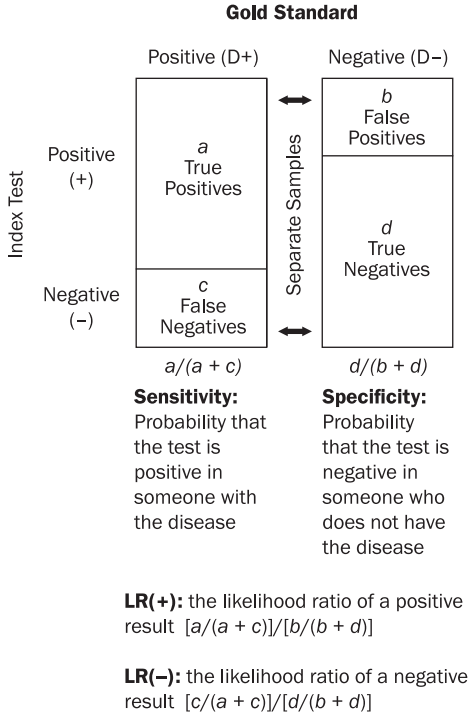
Bias is any systematic deviation of an estimate from the true value. As with all clinical research studies, test accuracy studies are susceptible to bias. A biased study of a dichotomous test will overestimate or underestimate the test's sensitivity, its specificity, or both. While several prior reviews discuss the identification of bias in test accuracy studies,[9–11] we describe the direction of a bias's effect on sensitivity and specificity and provide detailed numerical examples. Understanding not only the likelihood of a bias, but its probable direction, may permit useful inferences even from a flawed study of diagnostic test accuracy. We group biases in studies of test accuracy into five categories (Table 2): incorporation bias, partial verification bias, differential verification bias, imperfect gold standard bias, and spectrum bias. After discussing the importance of a study's sampling scheme, we will separately discuss each of these biases and the direction of its effect on sensitivity and specificity.

CASE–CONTROL VERSUS CROSS-SECTIONAL SAMPLING

Sampling for test accuracy studies is most often either case–control or cross-sectional (Figure 1). In a case–control design, D+ patients are sampled separately from D– patients. This allows measurement of sensitivity, specificity, LR(+), and LR(–), but it does not allow measurement of disease prevalence, PPV, or NPV because the "prevalence" is determined by the ratio of cases to controls set by the investigator. With cross-sectional sampling, the sample is defined by a characteristic such as clinical presentation that is separate from disease status or test result. This allows measurement of a meaningful prevalence (or pretest probability), PPVs, and NPVs in addition to sensitivity, specificity, LR(+), and LR(–). It is also possible to sample separately on index test result. For a dichotomous index test, patients with positive test results could be sampled separately from patients with negative test results. Such a study would allow calculation of PPVs and NPVs but not sensitivity, specificity, LR(+), LR(–), or prevalence. In practice, separating patients based on test result is usually done in the context of a cross-sectional study that provides the overall proportion of patients with each test result.

Some test accuracy studies fail to make their sampling schemes explicit. Equal numbers of D+ and D– patients suggest that they were sampled separately in a case–control design. However, some studies that use case–control sampling have different numbers of D+ and D– patients, conveying the impression of cross-sectional sampling. This is particularly misleading if the authors calculate and report PPVs and NPVs. For example, Davis et al.[12] did a case–control study to assess the accuracy of the "classic triad" of fever, spine pain, and neurologic deficit to predict spinal epidural abscess in ED patients. They identified 63 cases of spinal epidural abscess and matched two controls per case resulting in $2 \times 63 = 126$ controls. Controls were ED patients with

Case-Control Sampling: Patients with the disease in question (D+ patients) are sampled separately from patients without the disease in question (D– patients).



Cross-Sectional Sampling: Patients with and without the disease in question (D+ and D– patients) are sampled together from a population defined by a characteristic such as clinical presentation that is separate from disease status or test result.

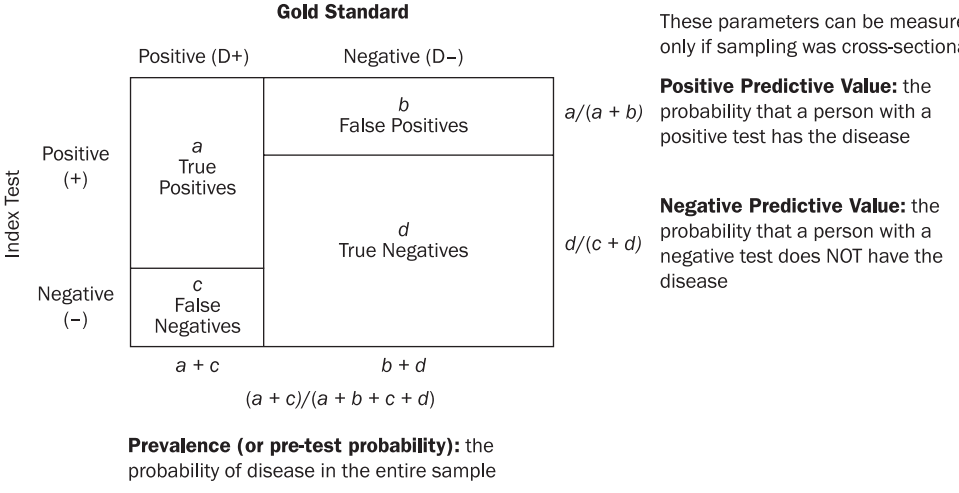


Figure 1. Dichotomous tests: definitions and case–control vs. cross-sectional sampling. Case–control sampling: patients with the disease in question (D+ patients) are sampled separately from patients without the disease in question (D– patients).

spine pain but without spinal epidural abscess. Since five of the 63 cases and one of the 126 controls had the classic triad, the authors reported a PPV of 5/6 = 83.3%. Readers could erroneously estimate the probability of spinal epidural abscess in a patient with the classic triad at 83.3%, but this is reasonable only if the prevalence of epidural abscess in similar patients is 1/3, since that was the "pretest probability" set by the authors when they chose two controls per case. The NPV of the classic triad was reported to be 68%, meaning that 100% – 68% = 32% of all ED patients with spine pain but *without* the classic triad have spinal epidural abscess.

In addition to precluding calculation of pretest probability and predictive values, case–control sampling often leads to unrepresentative samples of D+ and D– patients. This is discussed below under "Spectrum Bias."

Table 2
Biases in Studies of Diagnostic Test Accuracy

Bias Type	Definition	Most Common Direction of Bias: Sensitivity Is Falsely …	Most Common Direction of Bias: Specificity Is Falsely …	Questions to Ask
Incorporation bias (includes review bias)	Classification of disease status partly depends on the results of the index test. The gold standard *incorporates* the index test. If the gold standard is expert clinical review, this includes failure to blind the expert(s) to the results of the index test.	**Raised**	**Raised**	How were patients classified into the D+ and D– groups? If the gold standard was the result of a panel of tests, was the index test excluded from the panel? If the gold standard was an assessment of disease status by experts, were they blinded to the index test results?
Partial verification bias (a.k.a. verification bias, referral bias, ascertainment bias, work-up bias)	Patients with positive index tests are more likely to get the gold standard, and only patients who get the gold standard are included in the study.	**Raised**	**Lowered**	Did inclusion in the study require application of a single gold standard test? If so, was the decision to apply that gold standard independent of the index test result?
Differential verification bias (a.k.a. double gold standard bias)	Patients with a positive index test are more likely to receive an immediate, invasive gold standard, whereas patients with a negative index test are more likely to receive clinical follow-up for development of disease. Bias occurs only if there is a subgroup where the two gold standards give different answers.	**Raised** For disease that can resolve spontaneously. **Lowered** For disease that only becomes detectable during the follow-up period.*	**Raised** For disease that can resolve spontaneously. **Lowered** For disease that only becomes detectable during the follow-up period.*	Was an immediate gold standard applied preferentially to patients with a positive index test and clinical follow-up applied preferentially to patients with a negative index test? If so, will the clinical follow-up give the same answer as the immediate gold standard? For example, no patients would be positive on the immediate test but negative on clinical follow-up because of spontaneous resolution of the disease.
Imperfect gold standard bias (a.k.a. C\copper standard bias)	The standard which determines a patient's true disease status actually misclassifies some patients.	**Raised** If errors on the index test and copper standard are correlated (conditional on true disease status). **Lowered** If errors on the index test and the gold standard are independent (conditional on true disease status).	**Raised** If errors on the index test and copper standard are correlated (conditional on true disease status). **Lowered** If errors on the index test and the gold standard are independent (conditional on true disease status).	Does the gold standard always correctly classify the target condition?

Table 2
continued.

Bias Type	Definition	Most Common Direction of Bias: Sensitivity Is Falsely …	Most Common Direction of Bias: Specificity Is Falsely …	Questions to Ask
Spectrum bias, disease and nondisease	Spectrum of disease and nondisease differs from clinical practice. Sensitivity depends on spectrum of disease. Specificity depends on spectrum of nondisease or of diseases that might mimic the disease of interest.	**Raised** When disease is skewed toward higher severity than in clinical practice—"sickest of the sick."	**Raised** When nondisease is skewed toward greater health—"wellest of the well."	Were the D+ and D– groups sampled separately? If so, 1. Was the spectrum of disease appropriate? Were clinically mild cases of disease included in the D+ group? 2. Was the spectrum of nondisease appropriate? Did the D– group include an appropriately heterogeneous group of individuals who might realistically be suspected of having the target condition?
Spectrum bias, test result (exclusion of ambiguous test results)	Patients with ambiguous or intermediate test results are excluded.	**Raised** Excluded patients with disease were more likely than those included to have been called falsely negative.	**Raised** Excluded patients without disease were more likely than those included to have been called falsely positive.	Were patients with intermediate/ambiguous test results included in the study?

*Not discussed in text of article.

CATEGORIES AND DIRECTION OF BIAS IN STUDIES OF TEST ACCURACY

Incorporation Bias

For a study of a diagnostic test to be valid, a positive index test should have no role in determining whether the gold standard classifies the patient as D+. If the index test is incorporated into the gold standard, both sensitivity and specificity will be biased up relative to a study that uses an independent gold standard. The circularity should be apparent; how can you determine whether a test identifies D+ patients if you define a D+ patient as someone with a positive test? Nevertheless, a study reporting the accuracy of troponin T in predicting major cardiac events defined one major cardiac event (myocardial infarction) by an elevated troponin T.[13] In the authors' defense, the primary goal of the study was to evaluate the accuracy of myeloperoxidase, not troponin T. As another example, Mower[10] reviewed a study of the accuracy of radiolabeled biliary tract imaging that included the scan results in the criteria used to define cholecystitis.[14]

Sometimes the gold standard that determines disease status is review of clinical information by an expert or panel of experts. We include with incorporation bias the failure to blind the reviewers to the results of the index test, which is also referred to as review bias. In the study by Maisel et al.[15] of B-type natriuretic peptide (BNP) as a test for acute heart failure, the gold standard was the consensus diagnosis of two cardiologists who were blinded to the BNP and ED diagnosis, but who reviewed the patient's chart, including medical history, ECG, follow-up studies, and most significantly, chest x-ray. The chest x-ray was not the index test that the study was designed to evaluate, but the authors incidentally reported on its ability to predict heart failure. Because the reviewers who determined whether the patient had congestive failure were not blinded to the chest x-ray, it is not surprising that "the best clinical predictor of congestive heart failure was an increased heart size on chest roentgenogram."[15] This shows only that the heart size factored highly in the reviewing cardiologists' determination of acute heart failure.

Studies of test accuracy that use the treating physician's final diagnosis as the gold standard are subject to incorporation bias since the physician likely used the index test to help determine the final diagnosis. Studies of the accuracy of regional wall motion abnormalities on the ED echocardiogram for acute cardiac ischemia generally accept the clinician's diagnosis of "unstable angina" as the gold standard.[16,17] This means that the diagnostic accuracy (sensitivity and specificity) of the ED echocardiogram is overestimated, because its result undoubtedly contributed to the final diagnosis of acute ischemia versus another condition. A study of test accuracy that uses the clinician's diagnosis as the gold standard actually answers a different question: how well does the test result predict the diagnosis of disease? These studies are more about doctors' understanding of the disease than about the accuracy of the test.[18]

Partial Verification Bias

Partial verification bias (also known simply as verification bias,[19–21] referral bias,[22] ascertainment bias, or work-up bias[23–25]) occurs when patients who are positive on the index test are more likely to get the gold standard test and only those who get the gold standard are included in the study.[18,19,26] In a study of test accuracy, application of the gold standard (and consequent inclusion in the study) should not depend on the result of the index test.

This is a frequent problem when studying the accuracy of a sign or symptom for a particular disease. The presence of the finding makes further diagnostic work-up and definitive testing more likely. Limiting the study to patients who received definitive testing excludes some D+ and D– patients who were either falsely or truly negative for the finding. The exclusion of D+ patients who were falsely negative biases sensitivity up, while the exclusion of D– patients who were truly negative biases specificity down.

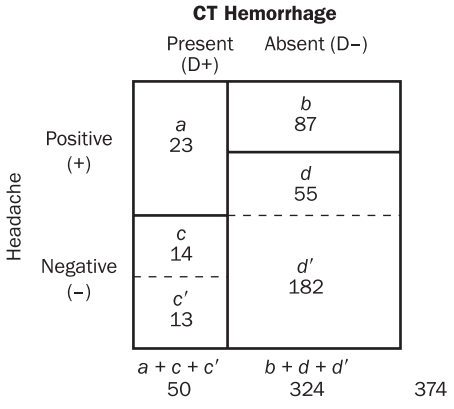
An extreme example is a study reporting on the accuracy of right lower quadrant (RLQ) pain for appendicitis in children.[27] This study found that RLQ pain was 96% sensitive and 5% specific for appendicitis. In other words, RLQ pain was present in 96% of the children with appendicitis and 95% of the children without appendicitis. The LR(+) was 1.0. Does this mean that RLQ pain is useless for diagnosing appendicitis in children? The gold standard for appendicitis in this study was the operative finding of an inflamed appendix, meaning that all 1,366 pediatric patients included in the study underwent appendectomies. The study actually shows that, to get an appendectomy, a child almost must have RLQ pain. Instead of limiting the study to patients who underwent appendectomies, a study could include all subjects evaluated for appendicitis and use clinical follow-up to determine disease status in those who do not go to the operating room. However, as discussed later, such a study is subject to bias if appendicitis sometimes resolves spontaneously.

As another example, consider the presence of headache as a test for hemorrhagic versus ischemic stroke in patients with acute neurologic deficit. In 2013, EPs commonly obtain a brain computed tomography (CT) scan, the gold standard to identify hemorrhagic versus ischemic stroke, in any patient with an acute neurologic deficit. In the 1980s CT scans were not obtained automatically. The presence of a headache often triggered the decision to order a CT because clinicians assumed that headache increased the likelihood of hemorrhage as a cause of the neurologic deficit. Panzer et al.[23] showed that, if CTs were more likely to be obtained in the evaluation of acute neurologic deficits when headache was present, then a study of headache as a predictor of hemorrhage would result in a higher sensitivity and lower specificity than if headache played no role in obtaining the gold standard CT (Figure 2). We follow Lijmer et al.[28] in using the term "partial verification bias" to distinguish it from "differential verification bias," which we previously called "double gold standard bias."[18]

Differential Verification Bias

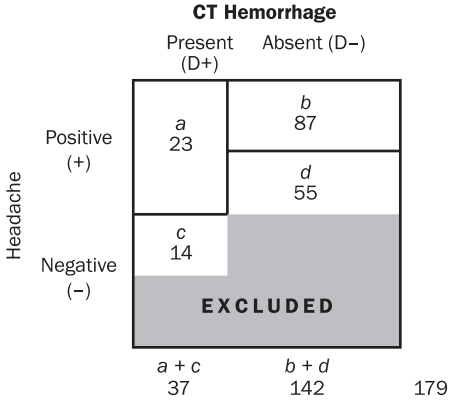
Differential verification bias[29] (also called double gold standard bias[18]) occurs when 1) different gold standards are used depending on the results of the index test and

All patients with acute neurologic deficits receive head CT scans, regardless of whether they have complaint or recent history of headache.



TRUE
Sensitivity: $a/(a + c + c')$ = 23/50 = 46%
Specificity: $(d + d')/(b + d + d')$ = 237/324 = 73%

Not all patients with acute neurologic deficits receive head CT scans, and a complaint or recent history of headache makes a CT more likely. (Only patients who received head CT scans are included in the study.)



BIASED
Sensitivity: $a/(a + c)$ = 23/37 = 62% INCREASED
Specificity: $d/(b + d)$ = 55/142 = 39% DECREASED

c' = patients with hemorrhage but no headache who would not have had a CT if presence of headache influenced the decision to order a CT. These false negatives would be excluded from a study where presence of headache influenced CT ordering and only patients with CTs were included.

d' = patients without hemorrhage who do not have headache and would not have had a CT if presence of headache influenced the decision to order a CT. These true negatives would be excluded from a study where headache influenced CT ordering and only patients with CTs were included.

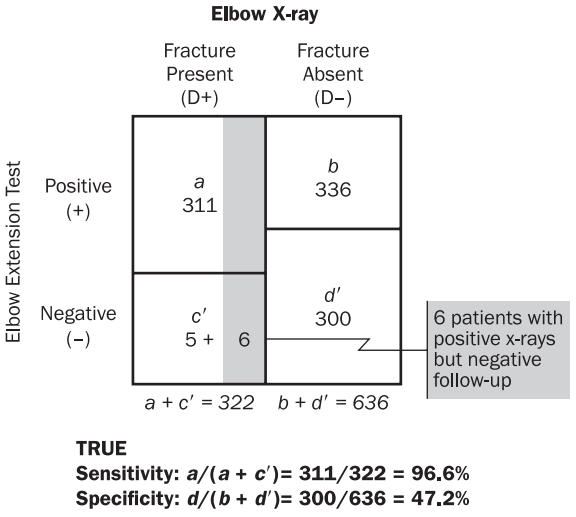
$c/c' > d/d'$ because patients with hemorrhage but without headache are more likely to have more severe neurologic deficits that prompt CT than patients without hemorrhage and without headache.

Figure 2. Verification bias in a study of headache as a predictor of intracranial hemorrhage in patients with acute neurologic deficits.[15]

2) the two gold standards can give different answers. As discussed earlier, partial verification bias can occur when the study excludes some patients with negative index tests because they did not receive the gold standard. Instead of excluding these patients, the investigators could follow them clinically to see if they develop the disease in question. As mentioned earlier, a study of RLQ pain as a predictor of appendicitis in children need not be limited to those who receive appendectomies. The investigators could include all children evaluated in the ED for appendicitis and use clinical follow-up to determine disease status in those who do not go to the operating room. This replaces the single gold standard (surgical diagnosis) with two gold standards—one (surgical diagnosis) applied preferentially to patients with positive index tests (RLQ pain) and one (clinical follow-up) applied preferentially to patients with negative index tests (no RLQ pain). In some cases, these two gold standards might give different answers. Cases of mild appendicitis might resolve spontaneously. If a child with mild appendicitis reports RLQ pain and therefore goes to the operating room, the inflamed appendix will result in a D+ classification and the presence of RLQ pain will be a true-positive. If the same child does not have RLQ pain and therefore gets clinical follow-up, the inflammation resolves spontaneously and the absence of RLQ pain is considered a true-negative. In this child with appendiceal inflammation that resolves spontaneously, the index test (RLQ pain) always gives the right answer. Sensitivity and specificity are artificially elevated relative to a study in which all patients receive the same gold standard independent of the index test result.

Appelboam et al.[30] studied the elbow extension test (inability to fully extend the elbow) as a predictor of elbow fracture in 958 adult ED patients. All 647 patients who had positive tests (were unable to extend fully) received x-rays (gold standard 1), but only 58 of the 311 patients with negative tests received x-rays (two of 58 = 3.5% showed fractures). The remaining 253 received clinical follow-up for subsequent elbow problems (gold standard 2, only three of 253 = 1.2% had problems on follow-up). Figure 3 provides a worked numerical example assuming that six (2.4%) of the 250 patients with negative index tests and negative follow-up would have had positive x-rays had they been done.[29] (The 58 extension test-negative patients who had x-rays were felt to be a random sample, and 3.5% had positive x-rays. Of the 253 additional extension

Single Gold Standard: All patients receive elbow x-rays regardless of elbow extension test result.



Double Gold Standard: Patients with positive elbow extension tests receive x-rays, but patients with negative elbow extension tests receive clinical follow-up. Some index-test negative patients who would have had positive x-rays and been counted as false negatives instead have negative clinical follow-up and are counted as true negatives.

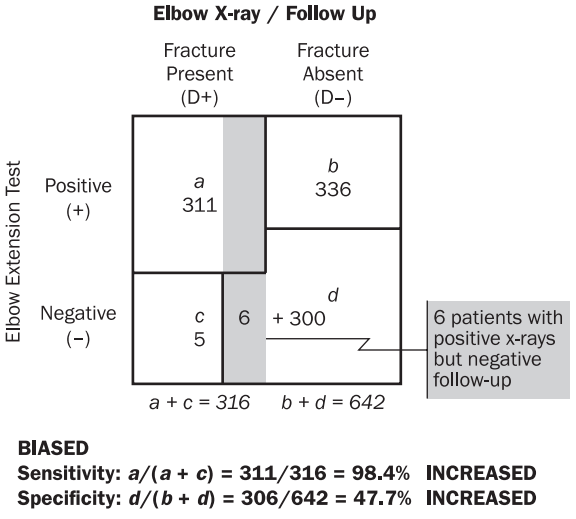


Figure 3. Double gold standard bias in a study of the elbow extension test as a predictor of elbow fracture.[4,22]

test-negative patients who did not get x-rays, only 1.3% had positive follow-up. Assuming based on the random sample that 3.5% would have had positive x-rays, then 3.5% – 1.3% = 2.2% had negative follow-up, but would have had positive x-rays; 2.2% × 253 = 6.) Under this assumption, the study overestimated both sensitivity and specificity.

In the original PIOPED study[31] to assess the accuracy of ventilation–perfusion scintigrams (V/Q scans) for pulmonary embolism (PE), 131 patients had normal or near normal scans. Only 55 of these 131 had the gold standard angiogram to diagnose PE. Five angiograms were positive. The other 76 patients were followed clinically and none developed PE. The authors were clear that 126 of 131, or 96%, represents an upper bound on the NPV of the V/Q scan, because some of the 76 patients who were followed clinically might have had positive angiograms.

The use of different gold standards depending on the index test result may be difficult to avoid for ethical or practical reasons. Apparently, the PIOPED investigators could not justify performing angiograms on all patients with normal V/Q scans. The degree to which sensitivity and specificity are distorted depends on 1) how frequently the index test result determines the choice of gold standard and 2) how often the two gold standards give different answers, which in turn depends on the natural history of the disease.

Imperfect Gold Standard Bias

A look at the periodic table of the elements suggests calling imperfect gold standard bias "silver" or "copper" standard bias depending on how imperfect the gold standard is. The effect of imperfect gold standard bias depends on whether the errors on the "copper standard" are correlated with errors on the index test. In the less common situation where the reasons for the errors are different, and misclassifications by the copper standard are independent of misclassifications by the index test (conditional on true disease status), both sensitivity and specificity of the index test will be falsely decreased.[32]

More often, errors on the index test and the copper standard will be correlated or "positively dependent."[33] For example, for a disease that occurs with a spectrum of severity, the D+ patients with mild or early disease who give false-negatives on the copper standard are also more likely to give false-negatives on the index test. This is referred to as "positive dependence for sensitivity" because a study comparing the index test to the imperfect gold standard generally results in falsely high sensitivity.[33] Similarly, given the heterogeneity of nondisease, the D– patients with conditions that resemble the target condition who give false-positives on the copper standard are also more likely to give false-positives on the index test. This is "positive dependence for specificity," because it generally results in falsely high specificity estimates.[33]

An example of imperfect gold standard bias with positive dependence for sensitivity arises when estimating the accuracy of a polymerase chain reaction (PCR) test for a specific infection such as *Bordetella pertussis*.[34,35] The traditional gold standard bacterial culture actually misclassifies some (D+) individuals with the infection as not having it, although culture rarely misclassifies a (D–) individual without the infection as having it. Relative to a hypothetical "true" gold standard, the culture has poor sensitivity but near perfect specificity; the problem is falsely negative cultures, not falsely positive cultures. In patients with genuine pertussis, the PCR is less likely to be falsely negative in a D+ patient with a (truly) positive culture than in D+ patients overall. Comparing the PCR to the imperfect gold standard culture will overstate sensitivity. Strebel et al.[34] did

pertussis serology, PCR, and culture on 212 patients with paroxysmal cough and determined that 27 had genuine pertussis, although only eight were positive on culture and only 10 were positive on PCR. While the true sensitivity of PCR was 10 of 27 or 37%, the apparent sensitivity compared with culture was five of eight, or 62.5% (Figure 4). Because the true sensitivity of the PCR (10 of 27 = 37%) was higher than the true sensitivity of the culture (eight of 27 = 30%), some true-positives (at least 10 – 8 = 2) on the PCR were counted as false-positives relative to the culture, and apparent specificity was falsely decreased. This study assumed on the basis of some good evidence that both culture and PCR were 100% specific.

Predicting the direction of imperfect gold standard bias's effect on sensitivity and specificity depends on whether the index test and the gold standard are positively dependent for sensitivity or specificity and requires additional assumptions such as the assumption of perfect specificity mentioned above.[33] Several approaches exist to adjust for imperfect gold standard bias, including using a composite "either-positive" gold standard or latent class analysis.[36–38] We agree with Begg[9] that the first step is to simply recognize when estimates of test accuracy come from a study that used an imperfect gold standard.

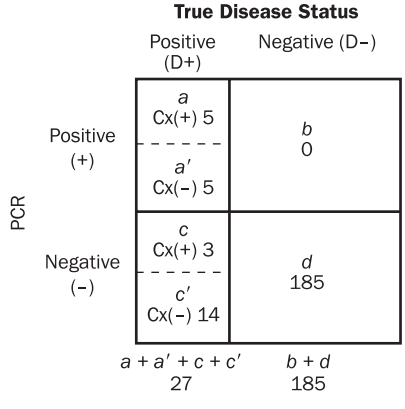
Spectrum Bias

Spectrum of Disease and Nondisease. As discussed earlier and in Figure 1, the PPV and NPV from a test accuracy study require cross-sectional sampling and depend on the prevalence of disease in the sample population. Sensitivity and specificity also depend on the sample population. More precisely, sensitivity depends on the spectrum of disease, and specificity depends on the spectrum of nondisease in the sample population.

We have been denoting patients with the disease as "D+," but some of them have mild or early disease, and others have severe or advanced disease; so instead of D+, we should have D+, D++, and D+++, and nondisease includes other diseases and varying states of health. A study with case–control sampling that limits D+ patients to the "sickest of the sick" will likely overestimate the sensitivity of the test when used clinically. If the study uses the "wellest of the well" as D– controls, it will likely overestimate the specificity. ED patients present with signs and symptoms of illness. If a test is for a particular disease, its specificity depends on what patients without that disease actually do have causing their signs and symptoms. These D– patients are often a much more heterogeneous group than the D– patients in a study of test accuracy that used case–control sampling.

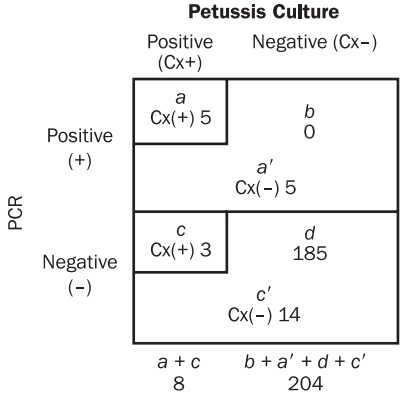
We previously mentioned the study by Maisel et al.[15] of the accuracy of BNP in diagnosing acute heart failure in ED patients with dyspnea. Among the 842 patients *without* acute heart failure were 72 patients who had chronic left ventricular dysfunction but whose acute shortness of breath was due to another cause such as bronchitis. These 72 patients had higher BNP levels than the other 770 patients without acute heart failure (mean = 346 pg/mL vs. 110 pg/mL). It appears that they were appropriately included in the D– population for purposes of calculating the specificity of 76% at a cutoff of 100 pg/mL.[39] In the ED, dyspnea patients who do *not* have acute heart failure but who do get BNP measured constitute a heterogeneous group including some patients with left ventricular dysfunction. Had these patients been omitted, the artificially homogeneous

True pertussis status in patients presenting with cough. Pertussis culture (Cx) is positive in only 8 out of 27 (29.6%) of patients with pertussis. Polymerase Chain Reaction (PCR) has slightly better sensitivity of 10/27 (37%). Neither Cx nor PCR is falsely positive in a patient without pertussis. PCR is less likely to be falsely negative when the Cx is positive (3/8 = 38%) than when culture is negative (14/19 = 74%) or overall (17/27 = 63%). This is positive dependence for sensitivity.



TRUE
Sensitivity: $(a + a')/(a + a' + c + c') = 10/27 = 37\%$
Specificity: $d/(b + d) = 185/185 = 100\%$

Apparent sensitivity and specificity of PCR relative to culture show a falsely high sensitivity and a falsely low specificity.



BIASED
Sensitivity: $a/(a + c) = 5/8 = 62.5\%$ INCREASED
Specificity: $(d + c')/(b + a' + d + c') = 199/204 = 97.5\%$ DECREASED

Figure 4. Imperfect gold standard bias in a study of PCR as a test for pertussis.[34]

D– group would have yielded a higher specificity of approximately 82% (Figure 5).

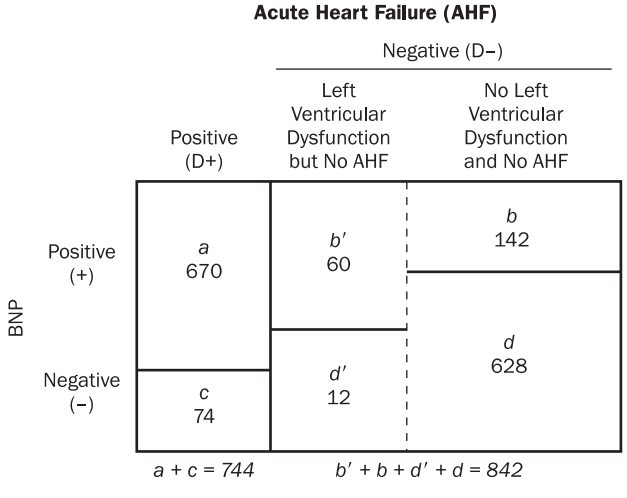
Spectrum of Test Results: Exclusion of Ambiguous Test Results. We include the bias resulting from exclusion of intermediate and ambiguous test results under the general heading of spectrum bias because this bias results from limiting the study to an unrepresentative spectrum of test results. In estimating the effect of excluding intermediate results, it is clearest to assume that they are no more likely in D+ patients than in D– patients; that is, they have an LR of 1. The effect of excluding intermediate results depends on how they would have been handled if included. Compared with treating intermediate results as *positive* for disease, excluding them biases sensitivity down and specificity up. Compared with treating intermediate results as *negative* for disease, the effect of excluding them is the opposite; sensitivity increases and specificity decreases. Rather than classifying all the intermediate results as either positive or negative, the study could require the test reader (if there is one) to make the choice between positive and negative for each individual patient. The effect of excluding intermediate results is then similar to spectrum bias that excludes D+ patients with mild disease and D– patients with severe nondisease (or other conditions that look like the disease). Sensitivity and specificity are both falsely increased. D+ patients with intermediate results have milder disease that the test reader is more likely to call falsely negative, and D– patients with intermediate results have more severe nondisease that the test reader is more likely to call falsely positive.

Consider the accuracy of the EP-performed bedside compression ultrasound (US) for diagnosis of deep vein thrombosis (DVT). Two studies[40,41] of a combined 146 patients showed perfect (100%) sensitivity and high specificity (pooled specificity = 95%) for the bedside US relative to a gold standard of color-flow duplex US performed by radiologists blinded to the bedside US result. Both of these studies used convenience sampling and may have excluded ambiguous results on the bedside US. A third similar study[42] of 183 patients showed much lower sensitivity (70%) and specificity (89%) relative to the same gold standard. This study used consecutive sampling and included patients with ambiguous US, forcing the bedside sonographer to state whether the examination was positive or negative. Assume that the convenience samples excluded patients with ambiguous US. Excluded patients with DVT might have been disproportionately false-negatives, and excluded patients without DVT might have been disproportionately false-positives. This could explain the higher accuracy in the convenience samples.

Sostman et al.[43] reported on the accuracy of V/Q scans done as part of the larger PIOPED II study[44] for diagnosing PE. Their calculations of sensitivity and specificity excluded intermediate test results. Figure 6 uses some of their data to provide a numerical example of how exclusion of intermediate results can falsely increase sensitivity and specificity.

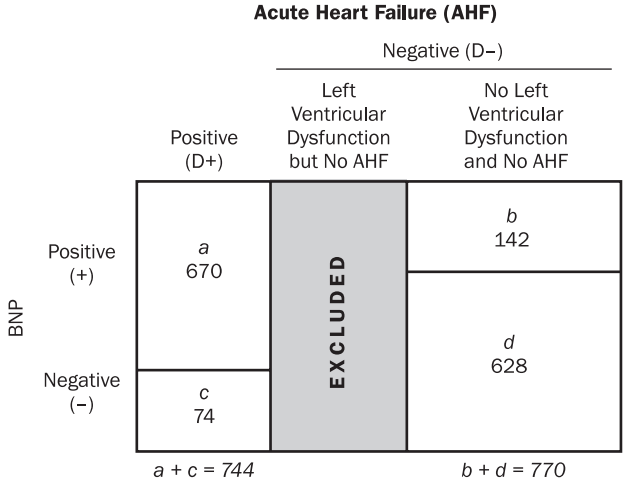
The best way to handle intermediate results is to report them, replacing the standard 2 × 2 table with a 3 × 2 table.[45,46] The index test is no longer dichotomous ("+" or "–"); it has three possible results ("+", "?", and "–"). The terms "sensitivity" and "specificity" do not have consistent interpretations for nondichotomous tests and should be avoided in favor of simply reporting the probability of each of the three results in the D+ and D– populations. The test now has three LRs: LR(+), LR (?), and LR(–), where "LR(?)" denotes the LR of an intermediate result. If LR(?) = 1, then excluding intermediate results does not change the numerical value of LR(+) and LR(–), but it does overstate the value of the test, and

ED dyspnea patients who do not have acute heart failure (AHF) include some patients with left ventricular dysfunction but other causes for their shortness of breath.



TRUE
Sensitivity: $a/(a + c)$ = 670/744 = 90%
Specificity: $(d' + d)/(b' + b + d' + d)$ = 640/842 = 76%

The effect of excluding the subgroup of patients with left ventricular dysfunction but no AHF creating an artificially homogeneous group of D– patients.



BIASED
Sensitivity: $a/(a + c)$ = 670/744 = 90% UNCHANGED
Specificity: $(d)/(b + d)$ = 628/770 = 82% INCREASED

Figure 5. Spectrum bias: heart failure (AHF) from a study of BNP as a test for AHF.[7,24]

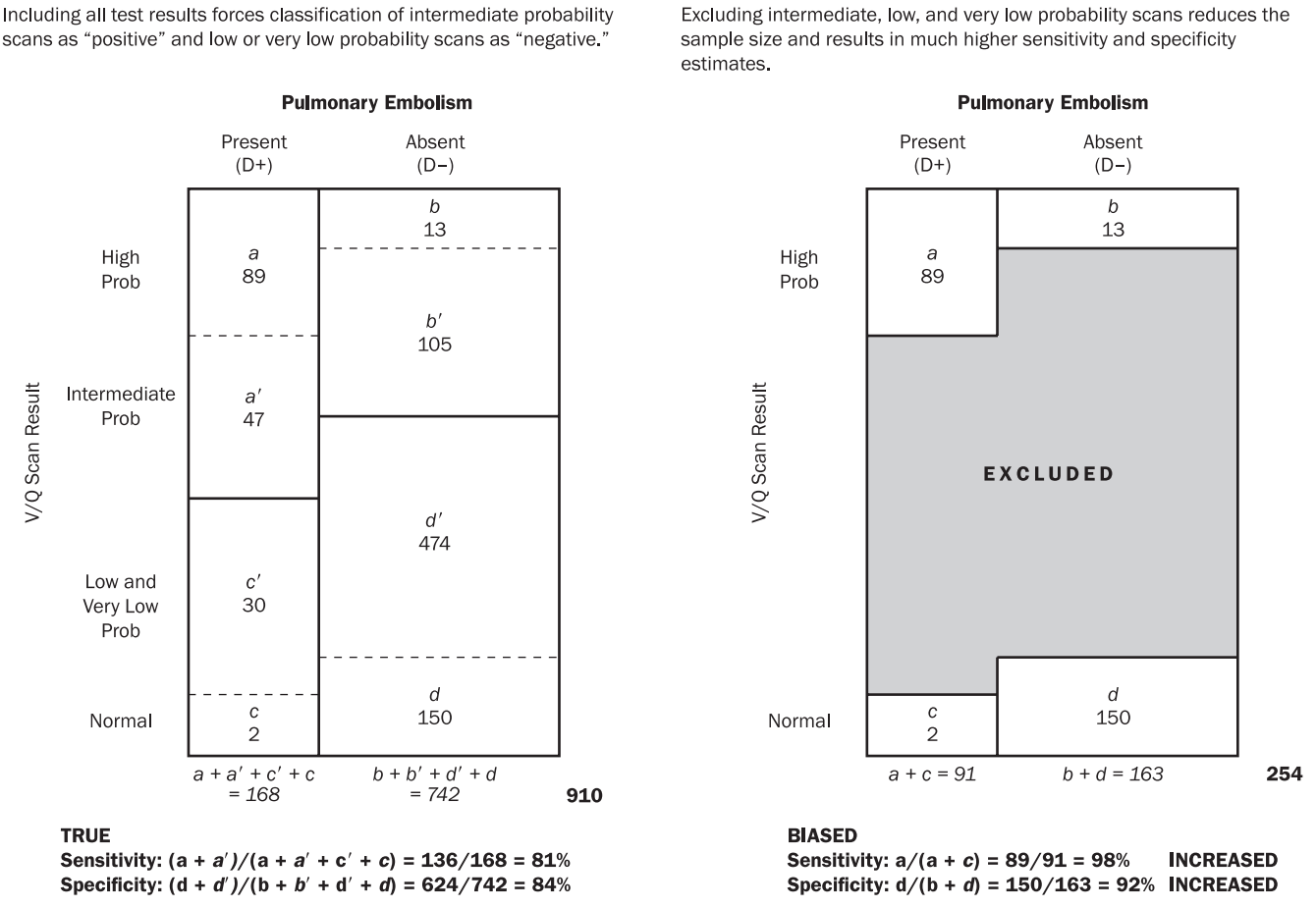


Figure 6. Effect of excluding intermediate test results in a study of V/Q scans for pulmonary embolism.

the test appears to be useful over too broad a range of pretest probabilities.

Making a Continuous Test Dichotomous. As our discussion of intermediate test results suggests, combining multiple distinct test results into only two categories, "positive" and "negative," discards useful information. Similarly, making a continuous test dichotomous by choosing a fixed cutoff to separate positive from negative reduces the value of the test. Instead of choosing a cutoff to make a continuous test dichotomous, we recommend dividing the continuous range of test results into three or more intervals and reporting interval LRs. This is discussed further elsewhere.[47,48]

Beyond Checklists

Several authors have proposed checklists of questions to determine whether a study of test accuracy is valid.[10,49,50] The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool is a 14-item checklist to help in the evaluation of diagnostic accuracy studies primarily for use in preparing systematic reviews.[51] Some of the items on the QUADAS list address the reliability (reproducibility) of the index test outside of the research setting. Reliability of diagnostic tests is beyond our current scope, but is addressed elsewhere.[52] The creators of the QUADAS checklist have released a more complex second version, QUADAS-2,[53] which has most of the same questions as the first version but broken into four domains: patient selection, index test, reference standard, and flow and timing. A checklist is useful in performing a systematic review when multiple reviewers are reading multiple test accuracy studies. It can also be useful in evaluating an individual study to identify potential problems. In Table 2, we have provided our own list of questions to help identify the potential biases discussed in this article. However, we encourage readers to go beyond identification of a potential bias and predict how it will affect the study results.

While the QUADAS and other checklists are intended for the readers of accuracy studies, the 25-item STARD (STAndards for the Reporting of Diagnostic accuracy studies) checklist is intended for the authors of accuracy studies.[54] By following this standard, authors provide readers with enough information to identify potential biases, as well as assess the reproducibility of the test and whether the results apply outside the study sample.

Just because a test accuracy study is subject to bias does not mean that it is useless in informing us about the test. If the biases in the study would inflate indices of test accuracy and they are still poor, we can conclude that the test is of limited usefulness. For example, a study of plain radiographs to identify malfunction of cerebrospinal shunts showed inadequate sensitivity.[55,56] This study used the neurosurgeon's final diagnosis of shunt malfunction as the gold standard and was subject to incorporation bias because the neurosurgeons used the radiographs to help make the diagnosis. Because

incorporation bias would give a falsely high estimate of sensitivity, we can still conclude that normal radiographs are inadequate to rule out shunt malfunction. Similarly, if a study uses an artificially homogeneous group of D– patients (such as healthy medical student volunteers) and still produces a low specificity estimate, we can conclude that the test will have a serious problem with false-positives in actual clinical practice where the D– patients will have other conditions that resemble the disease in question. The same study might use a realistic D+ group and produce an unbiased estimate of sensitivity. This was the case in the study of BNP for acute heart failure by Maisel et al.[15] mentioned earlier. A study such as the V/Q scan study by Sostman et al.[43] may exclude intermediate results in its summary report of test indices, but provide enough data to calculate the proportions with intermediate results and their LRs. This, in turn, allows accurate estimation of the test's value and the range of pretest probabilities for which the test is useful. In short, even flawed studies of diagnostic test accuracy can be useful.

CONCLUSIONS

The evaluation of test accuracy studies is important in emergency medicine. The reader of a test accuracy study should be able to 1) identify the index test being studied, the gold standard used to classify patients into D+ and D– groups, and the sampling scheme; 2) recognize the five categories of bias (incorporation bias, partial verification bias, differential verification bias, imperfect gold standard bias, and spectrum bias); and 3) predict (usually) the direction in which these biases will affect indices of test accuracy. Identifying potential biases and considering the direction in which they would distort the study results makes it possible to determine what useful information, if any, can be gleaned from the study.

References

1. Croskerry P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. Acad Emerg Med. 2002; 9:1184–204.
2. Lang ES, Worster A. Getting the evidence straight in emergency diagnostics. Acad Emerg Med. 2011; 18:797–9.
3. Kocher KE, Meurer WJ, Desmond JS, Nallamothu BK. Effect of testing and treatment on emergency department length of stay using a national database. Acad Emerg Med. 2012; 19:525–34.
4. Eddy D. Common Screening Tests. Philadelphia, PA: American College of Physicians, 1991.
5. Schunemann HJ, Oxman AD, Brozek J, et al. GRADE: assessing the quality of evidence for diagnostic recommendations. Evid Based Med. 2008; 13:162–3.
6. Gallagher EJ. Evidence-based emergency medicine/editorial. The problem with sensitivity and specificity … [comment]. Ann Emerg Med. 2003; 42:298–303.
7. Hayden SR, Brown MD. Likelihood ratio: a powerful tool for incorporating the results of a diagnostic test into clinical decisionmaking. Ann Emerg Med. 1999; 33:575–80.
8. Choi BC. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. Am J Epidemiol. 1998; 148:1127–32.
9. Begg CB. Biases in the assessment of diagnostic tests. Stat Med. 1987; 6:411–23.
10. Mower WR. Evaluating bias and variability in diagnostic test reports. Ann Emerg Med. 1999; 33:85–91.
11. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004; 140:189–202.
12. Davis DP, Wold RM, Patel RJ, et al. The clinical presentation and impact of diagnostic delays on emergency department patients with spinal epidural abscess. J Emerg Med. 2004; 26:285–91.
13. Brennan ML, Penn MS, Van Lente F, et al. Prognostic value of myeloperoxidase in patients with chest pain. N Engl J Med. 2003; 349:1595–604.
14. Eikman EA, Cameron JL, Colman M, Natarajan TK, Dugal P, Wagner HN Jr. A test for patency of the cystic duct in acute cholecystitis. Ann Intern Med. 1975; 82:318–22.
15. Maisel AS, Krishnaswamy P, Nowak RM, et al. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. N Engl J Med. 2002; 347:161–7.
16. Lau J, Ioannidis JP, Balk EM, et al. Diagnosing acute cardiac ischemia in the emergency department: a systematic review of the accuracy and clinical effect of current technologies. Ann Emerg Med. 2001; 37:453–60.
17. Lau J, Ioannidis JP, Balk E, et al. Evaluation of technologies for identifying acute cardiac ischemia in emergency departments. Evid Rep Technol Assess (Summ). 2000; 26:1–4.
18. Newman TB, Kohn MA. Critical appraisal of studies of diagnostic tests. In: Evidence-Based Diagnosis. New York, NY: Cambridge University Press, 2009, pp 94–115.
19. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983; 39:207–15.
20. Sox HC. The evaluation of diagnostic tests: principles, problems, and new developments. Annu Rev Med. 1996; 47:463–71.
21. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. Stat Methods Med Res. 1998; 7:337–53.
22. Geleijnse ML, Krenning BJ, van Dalen BM, et al. Factors affecting sensitivity and specificity of diagnostic testing: dobutamine stress echocardiography. J Am Soc Echocardiogr. 2009; 22:1199–208.
23. Panzer RJ, Suchman AL, Griner PF. Workup bias in prediction research. Med Decis Making. 1987; 7:115–9.
24. Philbrick JT. Work-up bias as an explanation for variations in sensitivity and specificity. J Gen Intern Med. 1989; 4:177.
25. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. J Clin Epidemiol. 1992; 45:581–6.

26. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. Biometrics. 2003; 59:163–71.
27. Pearl RH, Hale DA, Molloy M, Schutt DC, Jaques DP. Pediatric appendectomy. J Pediatr Surg. 1995; 30:173–8.
28. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999; 282:1061–6.
29. de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Bossuyt PM, Moons KG. Adjusting for differential-verification bias in diagnostic-accuracy studies: a Bayesian approach. Epidemiology. 2011; 22:234–41.
30. Appelboam A, Reuben AD, Benger JR, et al. Elbow extension test to rule out elbow fracture: multicentre, prospective validation and observational study of diagnostic accuracy in adults and children. BMJ. 2008; 337:a2428.
31. PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). The PIOPED Investigators. JAMA. 1990;263:2753–9.
32. Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. J Chronic Dis. 1981; 34:599–610.
33. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford, UK: Oxford University Press, 2003.
34. Strebel P, Nordin J, Edwards K, et al. Population-based incidence of pertussis among adolescents and adults, Minnesota, 1995-1996. J Infect Dis. 2001; 183:1353–9.
35. Baughman AL, Bisgard KM, Cortese MM, Thompson WW, Sanden GN, Strebel PM. Utility of composite reference standards and latent class analysis in evaluating the clinical accuracy of diagnostic tests for pertussis. Clin Vaccine Immunol. 2008; 15:106–14.
36. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009; 62:797–806.
37. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics. 1996; 52:797–810.
38. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol. 1995; 141:263–72.
39. Schwam E. B-type natriuretic peptide for diagnosis of heart failure in emergency department patients: a critical appraisal. Acad Emerg Med. 2004; 11:686–91.
40. Farahmand S, Farnia M, Shahriaran S, Khashayar P. The accuracy of limited B-mode compression technique in diagnosing deep venous thrombosis in lower extremities. Am J Emerg Med. 2011; 29:687–90.
41. Jang T, Docherty M, Aubin C, Polites G. Resident-performed compression ultrasonography for the detection of proximal deep vein thrombosis: fast and accurate. Acad Emerg Med. 2004; 11:319–22.
42. Kline JA, O'Malley PM, Tayal VS, Snead GR, Mitchell AM. Emergency clinician-performed compression ultrasonography for deep venous thrombosis of the lower extremity. Ann Emerg Med. 2008; 52:437–45.
43. Sostman HD, Stein PD, Gottschalk A, Matta F, Hull R, Goodman L. Acute pulmonary embolism: sensitivity and specificity of ventilation-perfusion scintigraphy in PIOPED II study. Radiology. 2008; 246:941–6.
44. Stein PD, Fowler SE, Goodman LR, et al. Multidetector computed tomography for acute pulmonary embolism. N Engl J Med. 2006; 354:2317–27.
45. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. Med Decis Making. 1987; 7:107–14.
46. Schuetz GM, Schlattmann P, Dewey M. Use of 3x2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. BMJ. 2012; 345:e6717.
47. Newman TB, Kohn MA. Multilevel and continuous tests. In: Evidence-Based Diagnosis. New York, NY: Cambridge University Press, 2009, pp 68–93.
48. Brown MD, Reeves MJ. Evidence-based emergency medicine/skills for evidence-based emergency care. Interval likelihood ratios: another advantage for the evidence-based diagnostician. Ann Emerg Med. 2003; 42:292–7.
49. Guyatt G, Rennie D, Meade MO, Cook DJ, eds. Users' Guide to the Medical Literature: A Manual for Evidence-Based Clinical Practice. Chicago, IL: AMA Press, 2002.
50. Straus SE, Richardson WS, Glasziou P, Haynes RB. Evidence-Based Medicine: How to Practice and Teach It. New York, NY: Elsevier/Churchill Livingstone, 2005.
51. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003; 3:25.
52. Newman TB, Kohn MA. Reliability and measurement error. In: Evidence-Based Diagnosis. New York, NY: Cambridge University Press, 2009, pp 10–38.
53. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011; 155:529–36.
54. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Br Med J. 2003; 326:41–4.
55. Mater A, Shroff M, Al-Farsi S, Drake J, Goldman RD. Test characteristics of neuroimaging in the emergency department evaluation of children for cerebrospinal fluid shunt malfunction. CJEM. 2008; 10:131–5.
56. Worster A, Carpenter C. Incorporation bias in studies of diagnostic tests: how to avoid being biased about bias. CJEM. 2008; 10:174–5.