Understanding Sources of Bias in Diagnostic Accuracy Studies

Robert L. Schmidt, MD, PhD, MBA; Rachel E. Factor, MD, MHS

● ***Context.***—**Accuracy is an important feature of any diagnostic test. There has been an increasing awareness of deficiencies in study design that can create bias in estimates of test accuracy. Many pathologists are unaware of these sources of bias.**

Objective.—**To explain the causes and increase awareness of several common types of bias that result from deficiencies in the design of diagnostic accuracy studies.**

Data Sources.—**We cite examples from the literature and provide calculations to illustrate the impact of study design features on estimates of diagnostic accuracy. In a companion article by Schmidt et al in this issue, we use these principles to evaluate diagnostic studies associated with a specific diagnostic test for risk of bias and reporting quality.**

Conclusions.—**There are several sources of bias that are unique to diagnostic accuracy studies. Because pathologists are both consumers and producers of such studies, it is important that they be aware of the risk of bias.**

(*Arch Pathol Lab Med.* 2013;137:558–565; doi: 10.5858/arpa.2012-0198-RA)

Accuracy is an important feature of any diagnostic test. Accuracy estimates play an important role in evidence-based medicine. They guide clinical decisions and are used to develop diagnostic algorithms and clinical guidelines. Poor estimates of accuracy can contribute to mistreatment, increased costs, or patient injury. Thus, it is important for accuracy estimates to be reliable.

There has been increasing awareness of deficiencies in study design and reporting in diagnostic test accuracy studies[1–5] and it is now recognized that diagnostic accuracy studies are subject to unique sources of bias. Pathologists are often involved in diagnostic accuracy studies and, as specialists in test methodology, play a key role in the generation of data on diagnostic accuracy. It is important for pathologists to understand the limitations of diagnostic studies and the methodologic issues that can lead to bias in accuracy estimates.

Over the years, there have been several efforts to make researchers aware of the methodologic issues associated with diagnostic tests. In 1999, the Cochrane Diagnostic and Screening Test Methods Working Group first convened to reduce deficiencies in diagnostic test reporting. Since then, the STARD (Standards of Reporting Diagnostic Accuracy) checklist,[6,7] QUADAS (Quality Assessment of Diagnostic Accuracy Studies) instrument,[8,9] and the QAREL (Quality Appraisal of Reliability Studies)[10] instrument have been introduced as evidence-based quality assessment tools to use in the systematic review of diagnostic accuracy studies. The STARD initiative alone has been adopted by more than 200 journals, spanning basic research to medicine. QUADAS has been widely adopted[11] and has been cited more than 500 times. It is recommended for use in systematic reviews of diagnostic accuracy by the Agency for Healthcare Research and Quality, Cochrane Collaboration, and the National Institute for Health and Clinical Evidence[12] in the United Kingdom.

See also p 566.

The problems associated with diagnostic tests are well recognized; however, the concepts involved are often subtle and unfamiliar to many pathologists. Because they play a key role in the production and interpretation of information on diagnostic test accuracy, it is important for pathologists to understand the types of bias that arise in diagnostic accuracy studies and their impact on accuracy estimates. Accuracy estimates are increasingly obtained from meta-analysis and, as noted above, an assessment of the risk of bias is now a standard part of any review of diagnostic accuracy. Owing to the increasing emphasis on evidence-based medicine, pathologists will be required to produce or interpret findings on the risk of bias in diagnostic studies.

Our objective is to provide an explanation of the common sources of bias in diagnostic studies. This information should help pathologists to identify risks of bias in diagnostic studies, to predict the impact of bias on study outcomes and, as producers of diagnostic studies, to avoid some of the methodologic issues that commonly cause bias in diagnostic studies.

Accepted for publication May 18, 2012.

From the Department of Pathology, University of Utah School of Medicine and ARUP Laboratories, Salt Lake City, Utah.

The authors have no relevant financial interest in the products or companies described in this article.

Reprints: Robert L. Schmidt, MD, PhD, MBA, Department of Pathology, University of Utah School of Medicine, 15 N Medical Dr E, Salt Lake City, UT 84112 (e-mail: Robert.schmidt@hsc.utah.edu).

FRAMEWORK FOR APPRAISAL

To be useful, a study must address a clinical question. Such questions are formulated in the familiar PICO format. For a diagnostic study, the PICO parameters are population, index test (the test under examination), comparator or reference test (the gold standard), and outcomes. The value of a study is a function of its capacity to resolve a clinical question. A clinical question can arise in the context of clinical work (Can this study help me to diagnose this patient's condition?), or in meta-analysis (Can this study help to resolve the question of the meta-analytic study?). To answer a clinical question correctly, a study must provide information that is both reliable (internal validity) and applicable (external validity). Internal validity is a function of bias and precision. A framework for appraisal is presented in Figure 1.

Bias is defined as a systematic difference in an observed measurement from the true value. For example, miscalibration causes systematic measurement errors that lead to analytic bias. In the context of a diagnostic accuracy study, bias occurs when the overall estimates of sensitivity or specificity systematically deviate from the real value. If bias exists, a study would *consistently* overestimate or underestimate the true accuracy parameters if the study were repeated. Thus, bias is error that does not "balance out" upon repetition. Unlike bias, precision is a function of random error and "balances out" upon repetition. Both bias and imprecision can render measurements unreliable. When measurements are unreliable, they may fail to represent the true value of the phenomenon being measured. A study is said to lack internal validity when it fails to measure what it purports to measure. Bias and random error (imprecision) are both sources of variation that can cause a measurement to differ from the true value. Both of these sources of variation negatively affect internal validity.

Bias and precision are a function of study design. Random error (imprecision) is determined by sample size and sound experimental design. Bias can occur at several different levels. In a diagnostic accuracy study it can arise from individual test measurements (analytic bias) or methodologic issues related to study design. Evaluation of internal validity involves an assessment of the risk of bias and the level of variability caused by imprecision. This, in turn, requires an assessment of study design features that could lead to bias or imprecision. Our discussion will focus on bias; however, it is important to recognize that internal validity is a function of both bias and imprecision. Risk of bias can only be evaluated if sufficient information about the study design is provided to allow for an assessment. Quality of reporting has a significant impact on internal validity which, in turn, is a determinant of study value.

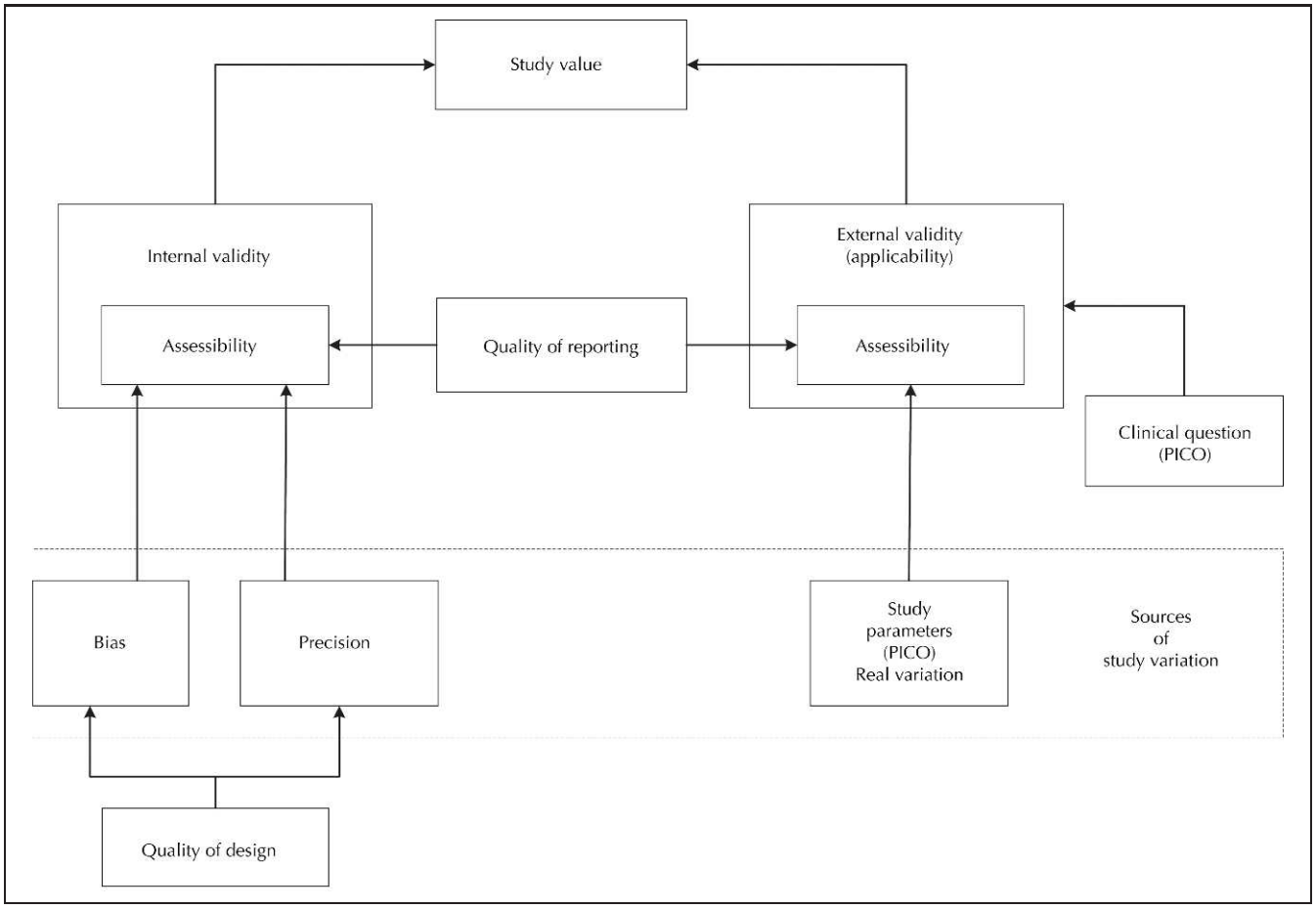

Figure 1. *Framework for critical appraisal of diagnostic accuracy studies. The figure provides a framework for critical appraisal of the value of a study. A study can only have value relative to a specific clinical problem. Value is determined by internal validity and external validity, which in turn depend on the ability to assess these factors. Assessment depends upon quality of reporting. Internal validity depends on bias and precision, which are determined by quality of design. Abbreviation: PICO, population, index test, comparator, outcome.*

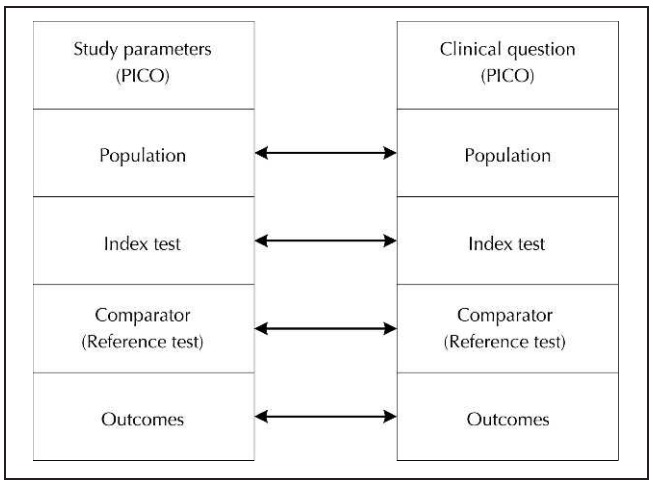

Figure 2. *Study applicability using the population, index test, comparator, outcome (PICO) framework. Assessment of applicability (external validity) depends on a point-by-point comparison of the components of the clinical problem: the population, index test, reference test, and outcomes.*

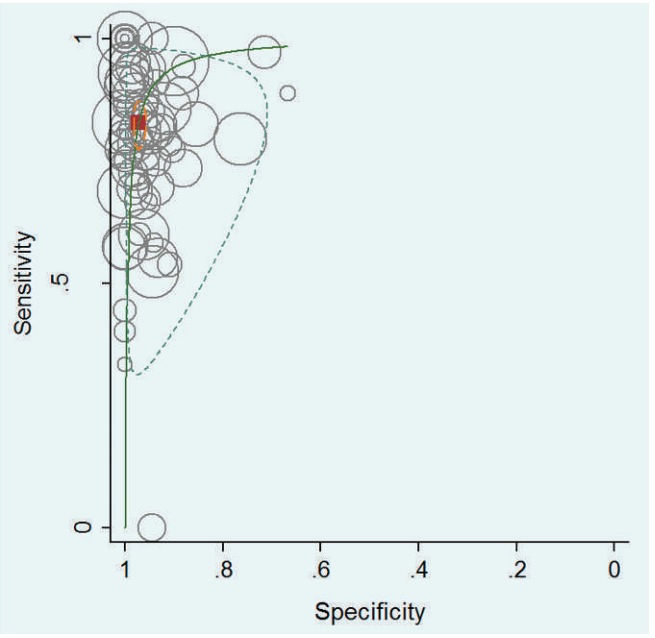

Figure 3. *Summary ROC curve showing heterogeneity in FNAC study accuracy. The figure shows a summary ROC curve for the diagnostic accuracy of FNAC for diagnosis of salivary gland lesions. Each circle represents a stud. The size of the circle is proportional to the weight given to the study in meta-analysis and each circle is centered at the point corresponding to the sensitivity and specificity of the study. The figure shows considerable variability in accuracy across studies. Abbreviations: FNAC, fine-needle aspiration cytology; ROC, receiver operating characteristic.*

The other determinant of study value is applicability. This is assessed by comparing the conditions of the study under evaluation (population, index test, reference test, outcomes) with those of the clinical question (Figure 2). Changes in any of these study parameters can cause changes in test accuracy. Such changes reflect true variability in test conditions and are not due to bias. For example, accuracy of fine-needle aspiration cytology (FNAC) might depend on the experience of the pathologist. The accuracy obtained in a study with an experienced cytologist would be higher than the accuracy obtained with a relatively inexperienced cytologist. If the difference were large, results from 2 studies conducted by pathologists with different levels of experience would necessitate a comparison of the different levels of experience. Because differences in methodology can lead to different outcomes, it is important for studies to fully report all of the methodology associated with both the index test and reference test so that sources of variation can be appreciated and applicability can be evaluated.

Since differences in methodology, patients, or other factors can lead to differences in accuracy measurements, studies conducted at different sites might show different levels of accuracy due to differences in the conditions at each site. Such differences cause difficulties in comparing studies, but this is again distinct from issues of bias.

Diagnostic studies often show considerable variation in outcomes. As discussed above, there are 3 possible reasons for variation: differences in study parameters, imprecision, and bias. As an example, Figure 3 depicts the results from a recent meta-analysis on the diagnostic accuracy of FNAC for parotid gland lesions. The results show considerable variability in accuracy and are quite heterogeneous. The heterogeneity implies that the studies differ owing to differences in study design (bias) or to real differences in the study parameters (PICO). Clearly, only a subset of these studies would be likely to provide valuable information with respect to a particular clinical question (eg, What is the accuracy of fine-needle aspiration [FNA] in a 1-cm lesion presenting in a US hospital, which appears benign on magnetic resonance imaging, was sampled with a 22-gauge needle with 4 passes, and was evaluated by a pathologist with 10 years of experience who specializes in head and neck tumors?). To make this determination, one would have to assess the reliability (risk of bias) and applicability of each study. This example provides a good example of the role of different types of variation in study appraisal.

In discussing issues of bias, we will refer to the QUADAS-2 framework.[12] QUADAS-2 is a survey that is used to assess the risk of bias in diagnostic studies and is organized into 4 domains: patient selection, index test, reference test, and patient flow. QUADAS-2 is closely aligned with the PICO format. QUADAS assesses risk of bias and applicability in each domain. Methodologic deficiencies can also give rise to subtle issues of applicability. We will discuss applicability that arises from design deficiencies but will not discuss applicability due to real differences in study conditions.

QUADAS-2 is designed to assess the value of a study with respect to a clinical question, but is not designed to assess reporting. The STARD guidelines are designed to insure high-quality reporting and can also be used to assess reporting quality. The assessment of reliability and applicability requires good reporting (Figure 1). QUADAS and STARD therefore serve 2 related but distinct functions.

PATIENT SELECTION

Differences in patient populations can affect accuracy, and the comparison of studies conducted in different populations raises questions of applicability.

Population Selection and Applicability

Study participants are obtained from a process of selection that starts from a target population and ends with the study participants (Figure 4). The study target population is

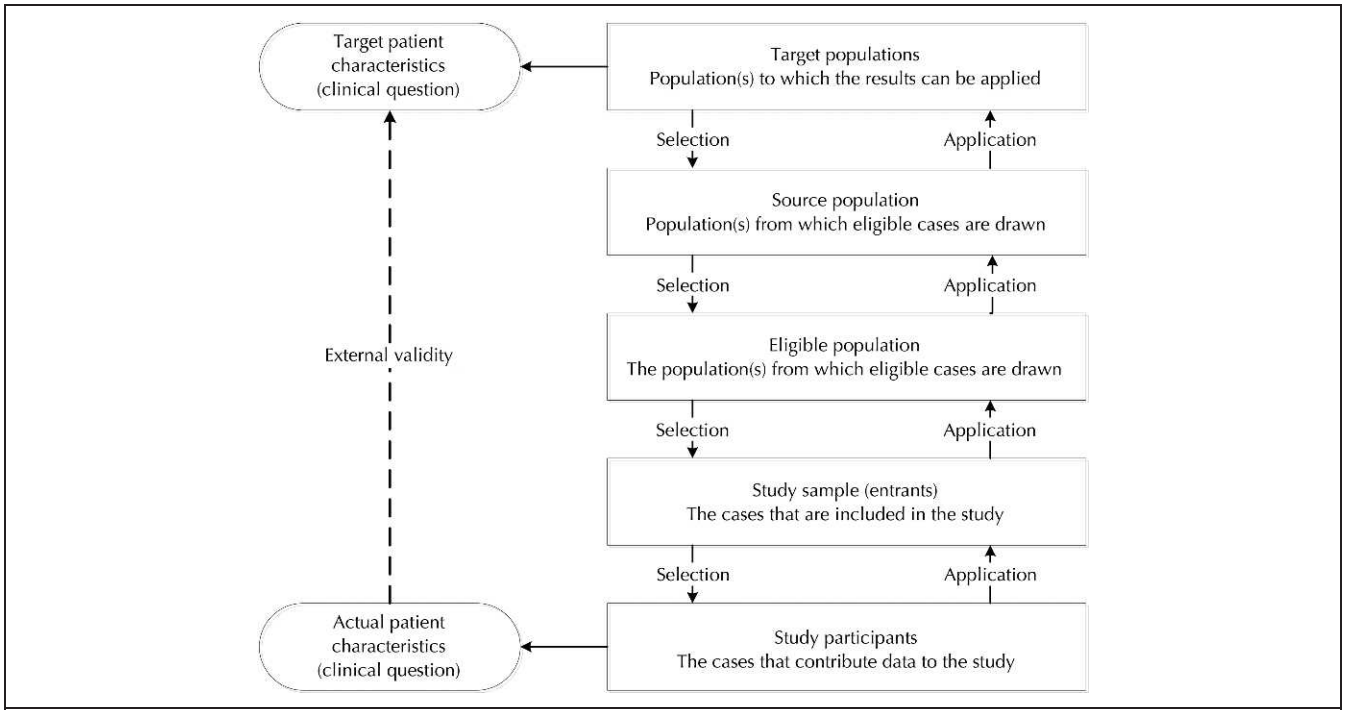

Figure 4. *Population concepts. The figure shows the relationship of several different populations that are used in population descriptions. The populations are related by a hierarchy of selection and application. Moving downwards, each population is selected from the population above. For example, the eligible population is obtained from the source population by the application of inclusion/exclusion criteria. Moving upwards, the results obtained with study participants are successively applied to each level, eventually reaching the target population. The results of the study can only be applied directly to the study participants, and applicability of the results obtained from the study participants must be successively inferred for each level to apply the results to a target population.*

conceptual and is obtained from a clinical problem (PICO). In a given study, the target population describes the patients for whom the results of the study are intended to apply. The extent to which the results of the study apply to the study target population depends on how well the actual study participants match the target population defined in the study question. The internal validity of a study depends on the applicability of the study participants to the study target population (ie, defined by the clinical question in the current study). Assessment of applicability requires an evaluation of each step of the selection process. For example, the study participants must be representative of the study entrants in order for the study participants. External validity depends on the applicability of the study participants to other target populations (ie, to clinical questions other than those posed by the present study) as shown in Figure 2.

Spectrum Bias

It is generally easier to detect advanced disease than early-stage disease, for which the signs are often subtle and difficult to distinguish from normal (Figure 5). The key parameter in patient spectrum is the *difference* in the measured test parameter between the disease and non-disease cases. We would expect diagnostic accuracy to be greater in a study conducted in a population with advanced disease than a population with less severe disease and, for this reason, studies may not be comparable if they are conducted on populations with significant differences in disease severity. Disease severity can be influenced by many factors such as the setting, referral patterns, and prior testing. All of these factors could give rise to differences in test performance that reflect actual differences in disease severity. Thus, it is important to fully describe the severity of disease in the patient population along with other factors that could be associated with disease severity. Although differences in disease severity are often referred to as "spectrum bias," we view these differences as issues of applicability because they reflect real differences in populations.

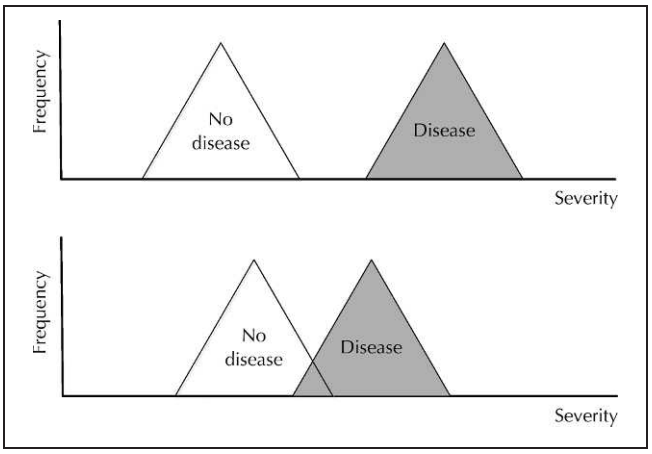

Figure 5. *Illustration of disease spectrum. The figure illustrates 2 studies that differ with respect to disease spectrum. In the upper panel, the patients with disease are widely separated from those without disease and one would expect high diagnostic accuracy in this situation. In the lower panel, the severity of disease shows overlap and diagnostic performance would be lower than in the upper panel.*

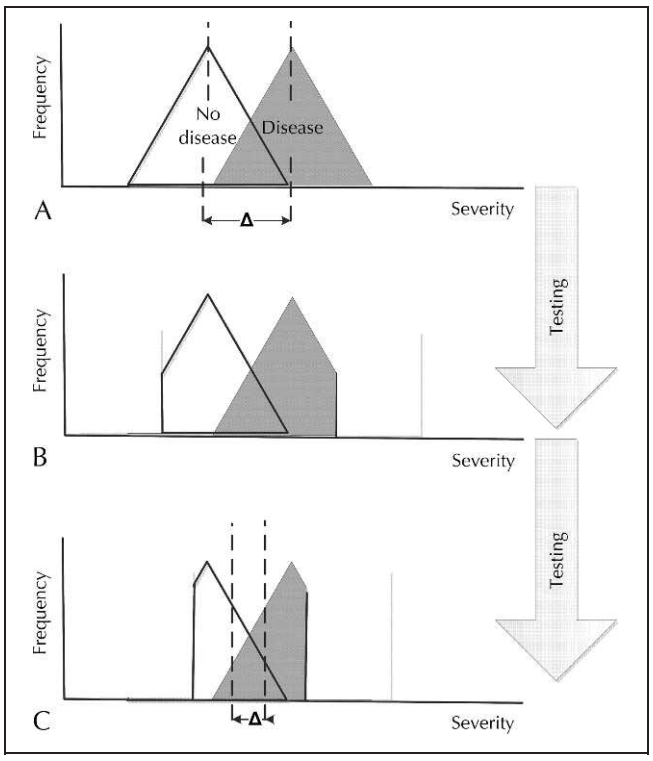

Figure 6. *Effect of referral pattern on patient spectrum. The figure illustrates how prior testing can change patient spectrum and complicate diagnosis. The initial distribution of diseased and nondiseased patients is shown in (A). The initial tests remove the easy-to-diagnose cases from the distribution, which creates the distributions in (B). Additional tests create the distributions in (C). The populations in (A) and (C) are not comparable because the patient spectrum is much narrower in (C) and test accuracy would be expected to be lower in (C) than in (A).*

Referral patterns are an important determinant of patient spectrum (Figure 6). Each stage in the referral process can produce diagnoses that remove cases from the initial distribution. Thus, the patient spectrum is altered by each referral. In general, one would expect the spectrum to narrow with each referral as "easy" cases are removed from the tails of the distribution. Test performance increases when the distribution is wide and, for that reason, diagnosis is more challenging at later stages than in the initial stages of the process, and a given test would be expected to be less accurate in later stages. Thus, prior testing and referral patterns can be an important factor when comparing test performance.

INDEX TEST

Diagnostic tests are often complex, multistep processes that can be performed in many different ways. For example, even for a simple procedure such as FNAC, there are many parameters involving the sample acquisition (needle size, number of passes, use of guidance techniques, experience of aspirator, use of rapid on-site evaluation, etc), sample processing (type of stain, use of ancillary techniques), and interpretation (number of pathologists who read the slide, experience level of the pathologist, availability of clinical information, etc). Each of these factors has the potential to affect test accuracy, and one can think of each variation as a different test with different performance characteristics. As indicated above, differences in accuracy that reflect differences in test conditions are not a source of bias but do give rise to issues of comparability. For example, is the accuracy of FNAC performed with a 22-gauge needle and ultrasound guidance equivalent to the accuracy obtained with a 26-gauge needle without guidance? Because differences in test conditions have the potential to affect results, it is important for studies to fully specify the methods.

We recently conducted a meta-analysis on the accuracy of FNAC for diagnosis of salivary gland lesions and found considerable heterogeneity in the results (Figure 3).[13] The question arose as to whether the variation in accuracy could be explained by differences in methodology. Unfortunately, the methods were insufficiently reported so that the effects of differences in methods could not be explored. In a subsequent study, we looked at the way in which FNAC studies described methods and found significant variation in reporting.[14] These examples illustrate the possible impact of test conditions on test results and why it is vital for studies to provide detailed descriptions of methods.

REFERENCE TEST

Classification Bias

No test is perfect. Errors in the reference test cause classification bias. There are 2 types of classification bias: differential misclassification and nondifferential misclassification. In differential misclassification, the error rate is associated with the index result. In the case of FNAC, positive FNA results may have a higher misclassification rate than negative FNA results, owing to error rates in histologic diagnosis. In nondifferential misclassification, the error rate is independent of the index test result, but this can underestimate sensitivity and specificity. An example is provided in Table 1. The magnitude of the bias depends on the disease prevalence, the accuracy of the index test, and the degree of misclassification. As shown in the example, misclassification can have significant effects. The misclassification rate can vary from site to site depending on the methodology associated with the reference test (eg, skill of the pathologist, use of ancillary tests). Misclassification can

Table 1. Effect of Classification Bias on Observed Sensitivity and Specificity[a]

	Perfect Reference Test		Imperfect Reference Test (10% Misclassification)	
	Reference Test		Reference Test	
Index Test	Positive	Negative	Positive	Negative
Positive	900	100	820	180
Negative	100	900	180	820
Total	**1000**	**1000**	**1000**	**1000**
Sensitivity	0.90		0.82	
Specificity	0.90		0.82	

[a] The left-hand column shows hypothetical results for an index test with 90% sensitivity and 90% specificity when evaluated by a perfect reference test. The right-hand column shows the same index test evaluated with an imperfect reference test. The imperfect reference test has a nondifferential misclassification rate of 10%. The number of true positives in the imperfect test is calculated as follows: True Positive = 900 (1 − 0.1) + 100 (0.1). The calculation shows that misclassifications cause observations to "move" across columns. Ten percent of the true positives are misclassified as false positives and 10% of the false positives are misclassified as true positives. Sensitivity decreases if the actual number of true positives is higher than the number of false positives.

be estimated by interrater reliability studies, but such studies are rarely referenced in FNAC diagnostic accuracy studies.

Diagnostic Review Bias and Incorporation Bias

These types of bias occur when the interpretation of the reference test is not independent of the index test, which weakens the results of retrospective studies.

Diagnostic review bias occurs when the pathologist interpreting the final histopathology is aware of the FNA result. This can affect results in that a pathologist might search more carefully for evidence of cancer if the FNA result is positive, and a strong FNA result might influence the interpretation of a borderline histologic result. Clinically, while it is important to use all information when making a diagnosis, the bias that results weakens studies of diagnostic accuracy. A rigorous study would require either reporting that the results are blinded, or that the cases were reviewed again to obtain a blinded diagnosis. In our experience, reporting of blinding in FNAC studies is quite poor.

In some cases, the result of the index test is explicitly used as a criterion for the reference test. Incorporation bias is best exemplified by clinical laboratory testing, specifically in the evaluation of β-D-glucan for diagnosis of invasive fungal infections. Invasive fungal infections are traditionally diagnosed by culture, imaging, and biopsy. β-D-Glucan is a blood-based test that offers an opportunity for an earlier diagnosis. By the European Organisation for Research and Treatment of Cancer criteria, the gold standard for invasive fungal infections includes a positive β-D-glucan test result.[15] In this case, the index test comprises part of the gold standard. In FNAC studies, incorporation bias occurs when a positive FNAC result is accepted as the gold standard as sometimes occurs in FNAC accuracy studies of the lung and mediastinum. While this criterion may be reasonable in clinical practice, it is a source of bias in diagnostic studies.

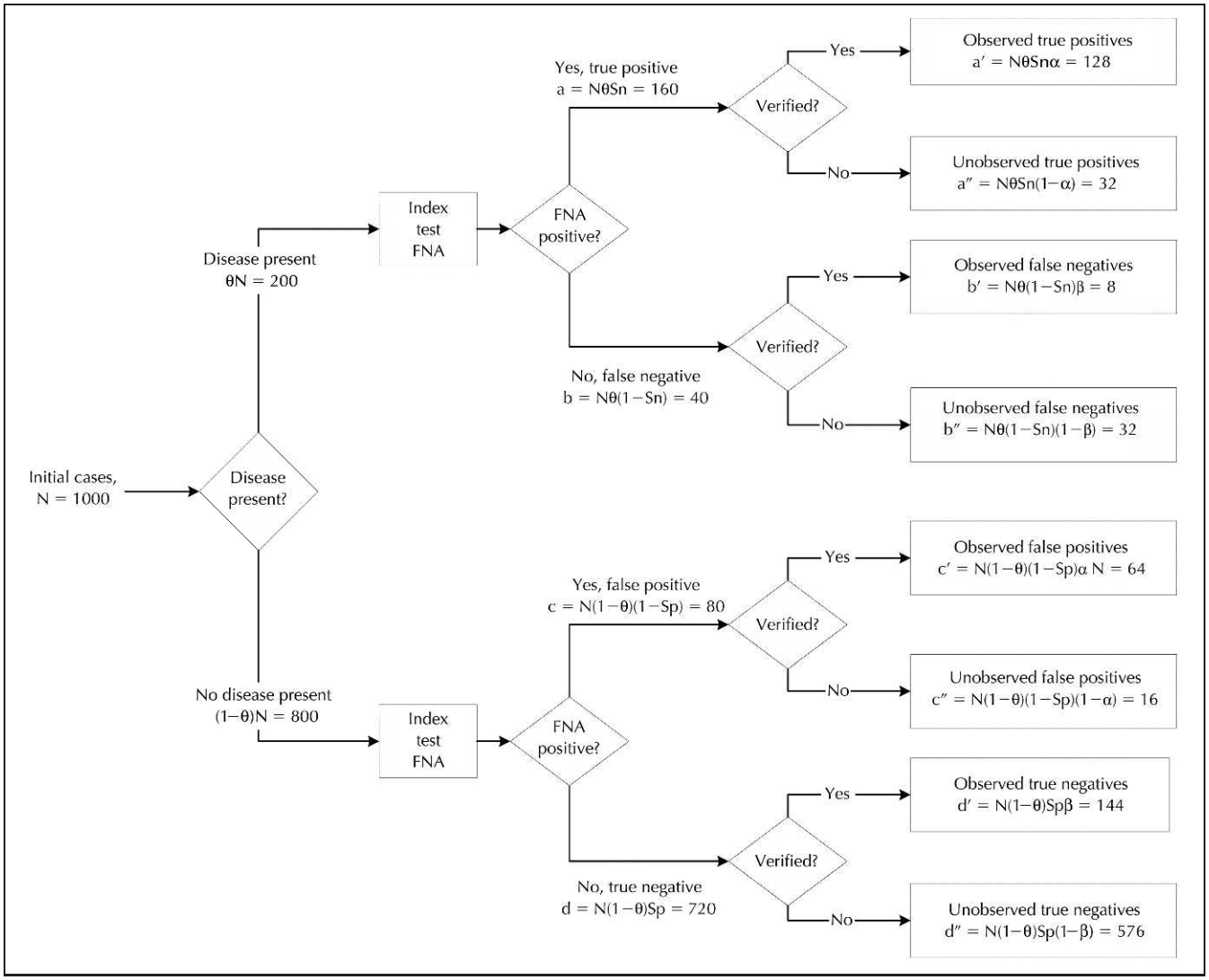

Figure 7. *Flow diagram for partial verification. The figure shows the effect of partial verification on the observed results of a diagnostic accuracy study. The assumptions are number of cases presenting for testing, N = 1000; disease prevalence, = 0.20; actual sensitivity (Sn) = 0.80; actual specificity (Sp) = 0.90; positive verification rate, α = 0.8; negative verification rate, β = 0.20. The flow diagram shows the number of cases in each category that will be observed. The bias in the observed accuracy statistics is shown in Table 3. Abbreviation: FNA, fine-needle aspiration.*

Table 2. Effect of Partial Verification on Accuracy Statistics[a]

	Actual		Observed	
	Reference Test		Reference Test	
Index Test	Disease	No Disease	Disease	No Disease
Positive	160	80	128	64
Negative	40	720	8	144
Total	**200**	**800**	**136**	**208**
Sensitivity	0.80		0.94	
Specificity	0.90		0.69	

[a] The table summarizes the results from the example in Figure 7. The observed results were obtained with partial verification and differ from the actual results (ie, the results that would have been obtained without partial verification). The example shows the bias that arises in observed results when partial verification is present.

PATIENT FLOW AND OUTCOMES

Partial Verification

Ideally, all those who are tested with the index test should receive verification by the reference test (gold standard). Failure to do so can cause bias in accuracy estimates and is known as partial verification bias. Partial verification can arise from different causes. A study may be designed so that positive cases are sampled more intensively than negative cases. Or a study may be designed so that all patients are referred for verification but, for various reasons, some patients do not present for verification. The first case represents a problem in design and the second represents a problem in study implementation. We discuss both types below.

Partial verification bias is common in FNAC accuracy studies for which the usual gold standard (histopathology) is invasive or expensive. Furthermore, most of these studies are retrospective, for which cases are identified from surgery or histopathology records. Such studies fail to record the results of those patients who received the index test but who did not receive surgery and histopathologic verification.

The example in Figure 7 demonstrates the effect of partial verification bias. In 1000 cases with a disease prevalence of 20%, there is an assumed sensitivity of 80% and specificity of 90%. Positive cases (ie, those with a positive FNA result) are verified at a higher rate (80%) than negative cases (20%). The results from the example are presented in Table 2, where the observed accuracy statistics are compared to the actual accuracy (ie, the accuracy that would be obtained with full verification). The table shows that sensitivity is falsely elevated from 80% to 94% and specificity is falsely decreased from 90% to 69%. In our experience, these numbers and the associated bias are typical for FNA studies.

It is important to note that partial verification only creates bias when the verification rate depends on the index test result. Partial verification would not occur if the positive and negative cases were randomly sampled at the same rate. Thus, if verification is limited by cost considerations, one can prevent bias by changing the sampling plan to make the verification rate independent of the outcome of the index test.

Withdrawals can have a similar impact if the withdrawal rate depends on the result of the index test. Withdrawals are common and occur for a variety of reasons. For example, patients initially screened at a community clinic may go to a tertiary care hospital for follow-up. Withdrawals can have the same effect as partial verification due to design; however, the magnitude of partial verification is generally less when it is due to withdrawals.

Differential Verification

Obviously, partial verification bias can be eliminated by verifying all cases; however, this would not be practical or ethical for invasive procedures. An alternative is to verify the remaining cases with a different reference test (a "brass standard") such as clinical follow-up. The problem with this solution is that the accuracy of the 2 reference standards may differ and the accuracy of the cases referred to the inferior test will suffer from classification bias. The overall accuracy estimates will be obtained from a combination of the biased and unbiased results. The resulting bias is called differential verification bias. To illustrate the effects of differential verification bias, we continue the FNA example from above but apply a different reference standard (eg, clinical follow-up) to the cases that were previously unobserved. We assume that the alternative brass standard has a 10% nondifferential misclassification rate. Differential verification can have a substantial effect as shown in Table 3.

These examples illustrate why documentation of flows is so critical in diagnostic accuracy studies. In our experience, withdrawals are often poorly documented in FNA diagnostic accuracy studies. The impact of partial verification bias can be estimated if the flows are well documented,[16] but it is better to prevent partial verification by good study design and management.

Inconclusive Results

Inconclusive results affect the applicability of one study to a population. Studies often aggregate test results or exclude

Table 3. Effect of Differential Verification[a]

	Observed Cases (Gold Standard)		Unobserved Cases		Observed Cases (Brass Standard)		Combined (Gold and Brass)	
	Reference Test		Reference Test		Reference Test		Reference Test	
Index Test	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Positive	128	8	32	16	30.4	17.6	158	26
Negative	8	144	32	576	86.4	521.6	94	666
Total	**136**	**152**	**64**	**592**	**116.8**	**539.2**	**252**	**692**
Sensitivity	0.94		0.50		0.26		0.63	
Specificity	0.95		0.97		0.97		0.96	

[a] The table is a continuation of the example in Figure 7 and Table 2. In the previous example, a significant number of cases were not verified (unobserved cases). The unobserved cases were mainly composed of fine-needle aspiration–negative cases. In this example, the unobserved cases are verified by an alternative reference standard (clinical follow-up). We assume that the alternative standard has a nondifferential misclassification rate of 10%.

results in particular categories. In FNA studies, common diagnostic categories include inadequate, negative for malignancy, atypical, suspicious, and positive for malignancy. As a first step, it is important that an article provide definitions for each of the indeterminate categories. Second, to maintain applicability, it is important that researchers report all results before aggregating results into categories. We often see articles in which results are grouped in different ways. For example, one article may count inadequate results and another article may exclude inadequate results from accuracy calculations. The different assumptions may be valid in the context of individual articles, but may not be applicable to other study populations.

It should be noted that the magnitude of the indeterminate rate can also affect applicability. The indeterminate rate can show significant variation between study sites. Differences in the indeterminate rate can reflect differences in criteria, differences in the sample population, or differences in methodology. Paradoxically, a study in which a cytopathologist defines 15% of cases as "indeterminate" may have better accuracy than a study with an indeterminate rate of 1% because the study with the high rate is only making a diagnosis on the easy cases.

SUMMARY

We have explained the basis of several common types of bias that are unique to diagnostic studies. Our objective has been to provide a framework to assist consumers of diagnostic accuracy studies in critically appraising results and to assist producers of diagnostic accuracy studies in avoiding many common sources of bias. It is important to recognize that no study is perfect and that bias and applicability are a matter of degree. Assessment of a study depends on quality of reporting. One cannot assess risk of bias or applicability of a study unless the details of the population, methods, and outcomes are fully reported. Thus, high-quality reporting is vital. These issues are likely to become more important in the future as evidence-based medicine increasingly relies upon systematic reviews and meta-analysis to study test performance.

References

1. Smidt N, Rutjes AWS, van der Windt DAWM, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology*. 2006;67(5):792–797.
2. Smidt N, Rutjes AWS, van der Windt DAWM, et al. Quality of reporting of diagnostic accuracy studies. *Radiology*. 2005;235(2):347–353.
3. Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol*. 2005;58(1):1–12.
4. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests [erratum in *JAMA*. 2000; 283(15):1963]. *JAMA*. 1999;282(11):1061–1066.
5. Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004;140(3):189–202.
6. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative—standards for reporting of diagnostic accuracy. *Clin Chem*. 2003;49(1):1–6.
7. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem*. 2003;49(1):7–18.
8. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, 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.
9. Whiting PF, Weswood ME, Rutjes AWS, Reitsma JB, Bossuyt PNM, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol*. 2006;6:9.
10. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol*. 2010; 63(8):854–861.
11. Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. *BMC Med Res Methodol*. 2011;11:27.
12. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011; 155(8):529–536.
13. Schmidt RL, Hall BJ, Wilson AR, Layfield LJ. A systematic review and meta-analysis of the diagnostic accuracy of fine-needle aspiration cytology for parotid gland lesions. *Am J Clin Pathol*. 2011;136(1):45–59.
14. Schmidt RL, Factor RE, Affolter KE, et al. Methods specification for diagnostic test accuracy studies in fine-needle aspiration cytology: a survey of reporting practice. *Am J Clin Pathol*. 2012;137(1):132–141.
15. De Pauw B, Walsh TJ, Donnelly JP, et al. Revised definitions of invasive fungal disease from the European Organization for Research and Treatment of Cancer/Invasive Fungal Infections Cooperative Group and the National Institute of Allergy and Infectious Diseases Mycoses Study Group (EORTC/MSG) Consensus Group. *Clin Infect Dis*. 2008;46(12):1813–1821.
16. Zhou X-H, Obuchowski N, McLish D. *Statistical Methods in Diagnostic Medicine*. 2nd ed. Hoboken, New Jersey: John Wiley and Sons; 2011.