Annals of Internal Medicine | Research and Reporting Methods

Evaluating Diagnostic Accuracy in the Face of Multiple Reference Standards

Christiana A. Naaktgeboren, MPH; Joris A.H. de Groot, PhD; Maarten van Smeden, MSc; Karel G.M. Moons, PhD; and Johannes B. Reitsma, MD, PhD

A universal challenge in studies that quantify the accuracy of diagnostic tests is establishing whether each participant has the disease of interest. Ideally, the same preferred reference standard would be used for all participants; however, for practical or ethical reasons, alternative reference standards that are often less accurate are frequently used instead. The use of different reference standards across participants in a single study is known as differential verification.

Differential verification can cause severely biased accuracy estimates of the test or model being studied. Many variations of differential verification exist, but not all introduce the same risk of bias. A risk-of-bias assessment requires detailed information about which participants receive which reference standards and an estimate of the accuracy of the alternative reference standard. This article classifies types of differential verification and explores how they can lead to bias. It also provides guidance on how to report results and assess the risk of bias when differential verification occurs and highlights potential ways to correct for the bias.

Ann Intern Med. 2013;159:195-202. www.annals.org

For author affiliations, see end of text.

A universal challenge in quantifying the accuracy of diagnostic tests or models is establishing whether each patient has the disease of interest (1). This classification is necessary to calculate various measures of diagnostic accuracy for the test being studied, such as sensitivity and specificity, predictive values, likelihood ratios, or receiver-operating characteristic curves (2). It is also a prerequisite for the derivation and validation of multivariate diagnostic prediction models. When making the classifications, researchers aim to use the best available method (the preferred reference standard) to verify the diagnosis in all participants.

Because of practical or ethical constraints, it is not always possible to ascertain disease status in all participants by using the preferred reference standard. Often an alternative, less accurate method (inferior reference standard) is used in patients who do not receive the preferred reference standard. The use of different reference standards in different groups of participants in a diagnostic study is known as differential verification (3). Differential verification is common; it may occur in up to one quarter of all diagnostic accuracy studies (4, 5).

As shown in **Table 1**, differential verification occurs for various reasons and in a wide range of clinical fields. A distinctive example comes from studies on the accuracy of mammography in detecting breast cancer (6, 11). The preferred reference standard, biopsy, is performed only when a lesion is found during mammography, indicating where to perform biopsy. In the ideal scenario, patients without lesions would also undergo this preferred reference standard, which would mean that they should have random biopsies. However, this option is not ethical and is not equivalent to a targeted biopsy. Instead, patients without lesions are followed to see whether breast cancer develops before the next screening.

Relying on disease classification that is based on an alternative reference standard may seem logical, but problems arise if one mistakenly treats the alternative and preferred reference standards as interchangeable when analyzing, reporting, or making inferences. When reference standards do not correspond well with the underlying "true" disease status (as is often the case with an inferior reference standard), the final disease classification will be wrong in some patients. The misclassification can cause biased estimates of diagnostic accuracy and regression parameters (12). When a mix of reference standards is used, index test or model accuracy estimates will systematically differ from the ideal study in which all patients have the preferred reference standard. This systematic deviation is called differential verification bias (4, 13, 14). **Figure 1** depicts a clinically relevant example of it.

This article elaborates on problems that can be confused with differential verification, proposes a classification system for types of differential verification, and explores the mechanisms by which each type leads to bias. It provides guidance on how to clearly report results when differential verification is present and how to assess and correct for the risk of bias. We believe that such guidance extends and improves STARD (Standards for the Reporting of Diagnostic Accuracy Studies) and QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) (13, 16).

Classification of Verification Patterns

Figure 2 gives an overview of the main verification patterns. Complete verification is the ideal situation in which all participants are verified by the same reference standard. A reference standard is often a single test, but when no single highly accurate test is available, multiple component tests may be used as a single reference method instead. Component tests can be used to make a final diagnosis in several ways. These include composite reference standards (a fixed rule for combining individual test results into a final diagnosis), panel diagnosis (consensus diagnosis by a group of experts), and latent class models (a statistical

Key Summary Points

Differential verification occurs in diagnostic accuracy studies when some patients receive a different, often less accurate, reference standard.

Differential verification may compromise the validity of study results if it is not properly analyzed and reported.

Clear reporting of differential verification requires addressing the accuracy of the alternative reference standard, providing information about why patients received the reference standards they did (the verification pattern), and analyzing and presenting results for each reference standard.

Judging the risk of bias due to differential verification requires considering the accuracy of the inferior reference standard and investigating the verification pattern.

method that uses associations among tests to define the unobserved disease status) (1). Although these methods use several pieces of information, the same information is available for all study participants and the same method is used to reach a final diagnosis.

The situation in which disease status is not ascertained at all in some patients, and is thus missing, is termed partial verification. When the missingness of the reference standard is somehow related to the index test results, index test accuracy estimates based on only the complete cases will be biased (17). Depending on the pattern of missingness, it may be possible to mathematically correct for partial verification bias (18, 19).

The situation in which different reference standards are used in different groups of patients is termed differential verification. The word *differential* means "varying according to circumstances or relevant factors" (20). A salient question in the case of differential verification is, "What factors determine which reference standard is used?" **Figure 3** shows 2 basic differential verification patterns. The key distinction between the patterns is the reason that patients received one reference standard over another.

Differential verification can be thought of as a missing data problem in which the preferred reference standard is missing in some patients and replaced by an alternative, inferior reference standard (21). Data can be missing completely at random, or the missingness can be related to measured or unmeasured patient characteristics (17, 22). If missingness is completely at random, estimates will be unbiased but inefficient, leading to CIs that will be wider than those obtained with complete verification. If missingness is dependent on patient characteristics, estimates are probably biased but the bias can be corrected for if these patient characteristics are measured.

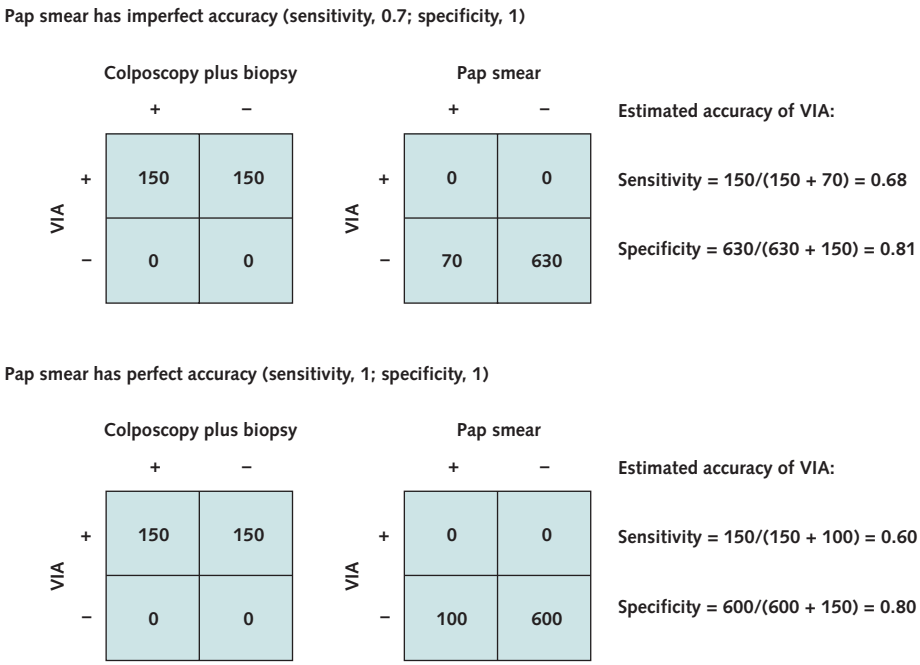
In pattern A depicted in **Figure 3**, reference standard assignment depends entirely on the index test results. The earliest definition of differential verification was limited to pattern A and was described as the situation in which "negative [index] test results are verified by a different, often less thorough, [reference] standard, for example follow-up" (4). This particular situation, in which all positive index test results are verified by the preferred reference standard whereas all negative index test results are verified by an inferior standard, was later termed *complete differential verification* (11), but we propose referring to it as *complete index test–dependent differential verification* for clarity. Pattern A either is a product of the clinical situation or is determined a priori by the researcher.

Frequently in diagnostic studies, the decision about which reference standard each patient receives is less straightforward and is determined by various factors besides, or in addition to, index test results. In pattern B depicted in **Figure 3**, the choice of reference standard is influenced by other factors, such as signs and symptoms, other test results, or patient or physician preference. Pattern B may be broken down into 3 subtypes based on the

Table 1. **Examples of Differential Verification**

Target Condition	Index Test	Preferred Reference Standard	Inferior Reference Standard	Reason for Not Using Preferred Reference Standard for All Patients	Reference
Breast cancer	Mammography	Biopsy	Follow-up	Impossible to perform a biopsy when no lesion is detected during mammography	6
Deep venous thrombosis	Diagnostic rule for ruling out deep venous thrombosis	Ultrasonography together with follow-up	Follow-up	To test the safety of the rule in clinical practice, patients with a normal score on the diagnostic rule were sent home without extensive testing	7
Congenital heart defect	Pulse oxometry (blood oxygen levels)	Clinical work-up	Congenital anomalies registry	Too expensive to perform a clinical work-up of all infants to detect a rare disease	8
Depression	Diagnostic prediction rule for depression	In-person interview	Telephone interview	Not practically feasible to interview all participants in person	9
Tuberculosis	Screening test	Sputum culture	Chest radiography and/or follow-up	Clinical records were used instead of setting up a diagnostic study	10

Figure 1. **An example of bias due to differential verification.**



The example is loosely inspired by a study on the accuracy of VIA in screening for cervical cancer (15). The preferred reference standard is colposcopy plus biopsy when a lesion is detected. Because the preferred standard is invasive, one might use an alternative, less invasive reference standard, the Pap smear, for participants with a normal VIA result. If one assumed that the Pap smear had perfect accuracy, the naive (biased) estimates of sensitivity and specificity for the VIA would be 0.68 and 0.81, respectively. If one recognized the sensitivity of the Pap smear as only 0.70, the true estimate of the sensitivity for the VIA would be 0.60. Pap = Papanicolaou; VIA = visual inspection using acetic acid.

pattern of "missingness" of the preferred reference standard (17). First, the choice of reference standard sometimes depends solely on known factors that are measured and recorded in the study. This pattern may occur, for example, when different study centers use slightly different imaging techniques as the reference standard because of availability of the technology. Second, unknown factors (those not recorded) could also influence this choice. An example of such a factor is patient preference: When the preferred reference standard is burdensome, some participants, particularly those with less severe symptoms, may opt out and be followed instead. Third, studies can be designed in

Figure 2. **Main verification patterns and terminology.**

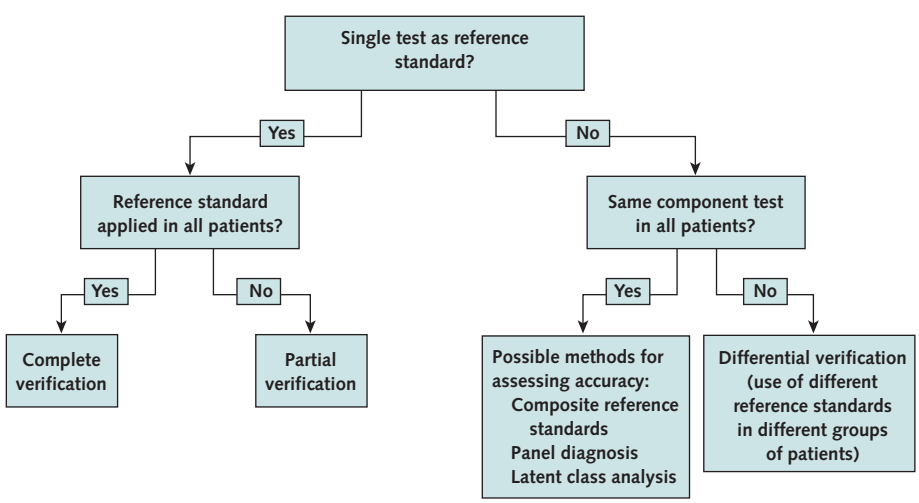
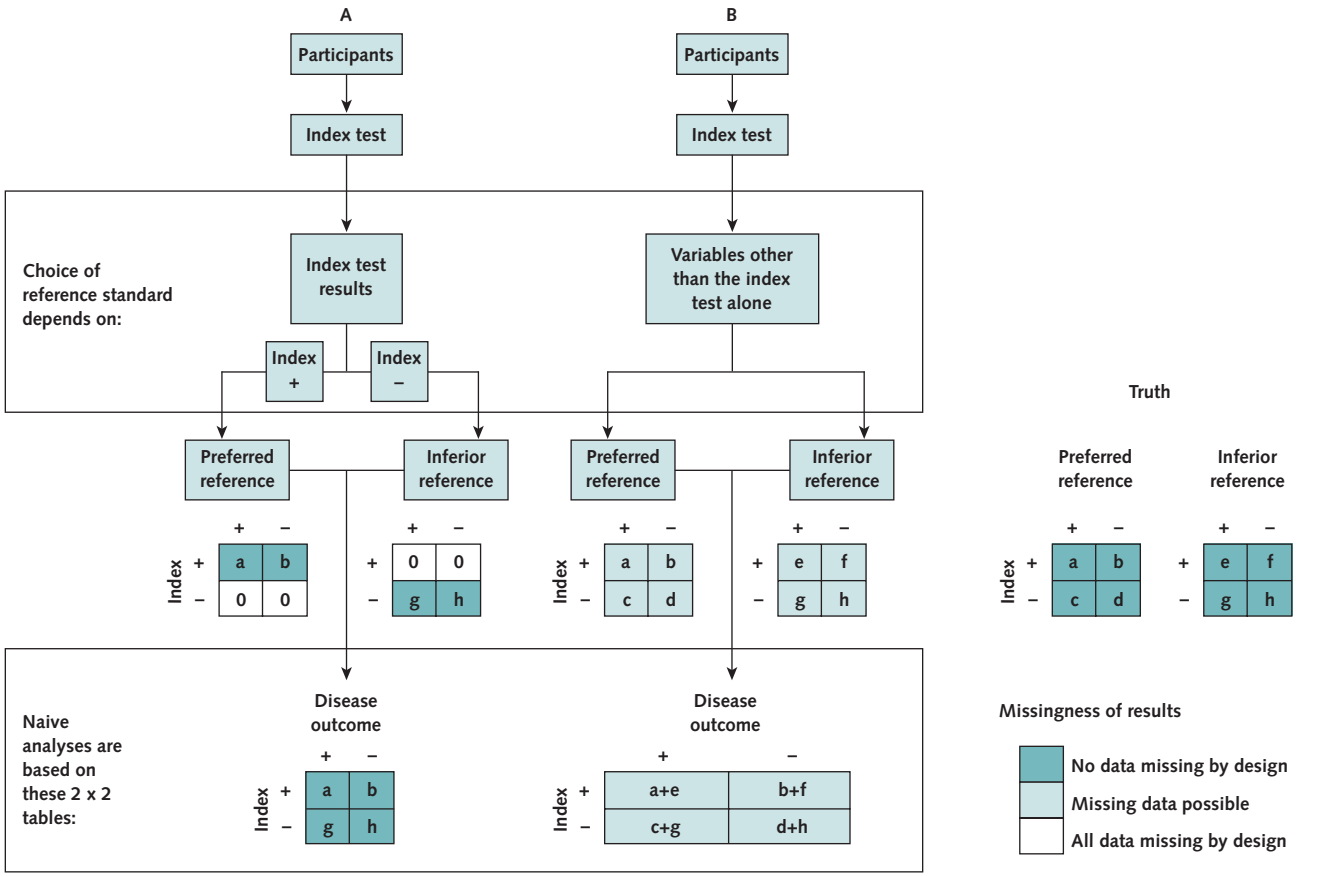


Figure 3. **Classification of differential verification patterns.**



In pattern A, the test a participant receives depends solely on the index test result, whereas in pattern B, other variables may influence reference standard assignment.

which the reference standard is assigned completely at random.

Factors Leading to Differential Verification Bias

The differential verification pattern—and, notably, the reasons that patients receive the reference standards they do—plays a major role in whether bias is introduced as well as whether this bias can be (partially) adjusted for. Current guidelines for assessing the risk of bias in a diagnostic study simply warn that there is a risk of bias when multiple reference standards are used (13). This guidance is supported by the results of 2 meta-analyses on factors influencing diagnostic accuracy estimates, which found that studies in which differential verification was present had a diagnostic odds ratio that was, on average, 1.6 (95% CI, 0.9 to 2.9) to 2.2 (CI, 1.5 to 3.3) times higher than that in studies of the same test that used a single reference standard (4, 5). The example in **Figure 1** illustrates how differential verification can lead to overestimates of accuracy. Although differential verification seems to generally result in a substantial overestimate of index test accuracy, it does not lead to clinically relevant bias in some situations.

Given that differential verification often causes bias, the pertinent question is whether the bias is large enough to be clinically relevant. The magnitude and direction of differential verification bias are influenced by various factors: the verification pattern, the accuracy of the reference standards, the proportion verified by each reference standard, and disease prevalence (11). Formal correction methods may address these factors, but it is difficult for the reader to consider multiple factors simultaneously when performing an explicit assessment of the risk of clinically relevant bias.

We recommend that readers break down the problem into a few questions that can independently be used to rule out the risk of differential verification bias (**Table 2**). Information about the disease prevalence and the proportion verified by each reference standard may be the most readily available information, but knowledge about either of these factors alone does not allow one to rule out the risk of clinically relevant bias. In many situations, the preferred reference standard is assumed—correctly or incorrectly—to have near-perfect accuracy. We recommend, therefore, that readers focus on the accuracy of the inferior ref-

erence standard and the verification pattern when ruling out the risk of clinically relevant differential verification bias.

Accuracy of the Inferior Reference Standard

As a general rule, the risk of clinically relevant differential verification bias decreases as the accuracy of the inferior reference standard increases. To estimate the accuracy of the inferior reference standard, its performance needs to be compared with that of the preferred reference standard. Although this may not be possible or ethical to do within the study, accuracy estimates may be available in the literature.

A commonly used alternative reference standard for making a final disease classification is follow-up (following patients over time to see whether symptoms worsen or improve). The follow-up information is used to decide retrospectively whether patients did indeed have the disease at the time the index test was done. Assessing the accuracy of follow-up is difficult, even when the preferred reference standard and follow-up are both done in a random subset of patients, because the condition of the patients can improve or worsen between when the preferred reference standard is performed and the end of follow-up.

Because estimating the accuracy of follow-up is difficult, it is particularly important to consider the quality and length of follow-up from a biological perspective, taking into account the natural course of existing cases as well as the incidence of new cases. We provide a cursory example of how this can be done using a study that investigated the accuracy of blood oxygen concentration measured at birth in detecting congenital heart defects (see **Table 1**) (23). In this study, newborns with low oxygen levels were treated by a cardiologist, whereas the rest were followed up after 1 year through a congenital anomalies register.

Follow-up should be long enough to allow as many hidden cases of disease to progress to a detectable stage as possible. However, if follow-up is too long, new cases developing after the index test was performed will also be detected. In the example, follow-up might be considered too short because some types of congenital heart defects are detected later in life. On the other hand, it was not too long because congenital heart disease is, by definition, already present at birth. The second point to consider is whether follow-up allows detection of the same type and severity of disease as the preferred reference standard. Researchers should focus on whether the test being studied detects cases that will benefit from clinical intervention rather than simply the presence of any disease (24). In the example, more serious types of defects are probably detected at birth, whereas less serious ones are detected during follow-up. Although this may not be a problem, if follow-up instead detects the more pronounced cases, the estimated sensitivity of the index test will be an overestimate of its sensitivity in detecting serious cases.

When the inferior reference standard is believed to have high accuracy, clinically relevant differential verification bias is unlikely and there is no need to look into the other factors influencing bias. This was believed to be the case in an example of differential verification from a study involving a clinical prediction rule for screening for depression in primary care (see **Table 1**) (9). In this study, in-person or telephonic questionnaires were used as the reference standard, but because the authors had reason to assume that these methods for assessing depression had similar accuracy, they argued that clinically relevant differential verification bias was unlikely. When the inferior reference standard's accuracy is questionable, however, the next step is to consider the verification pattern.

Verification Pattern

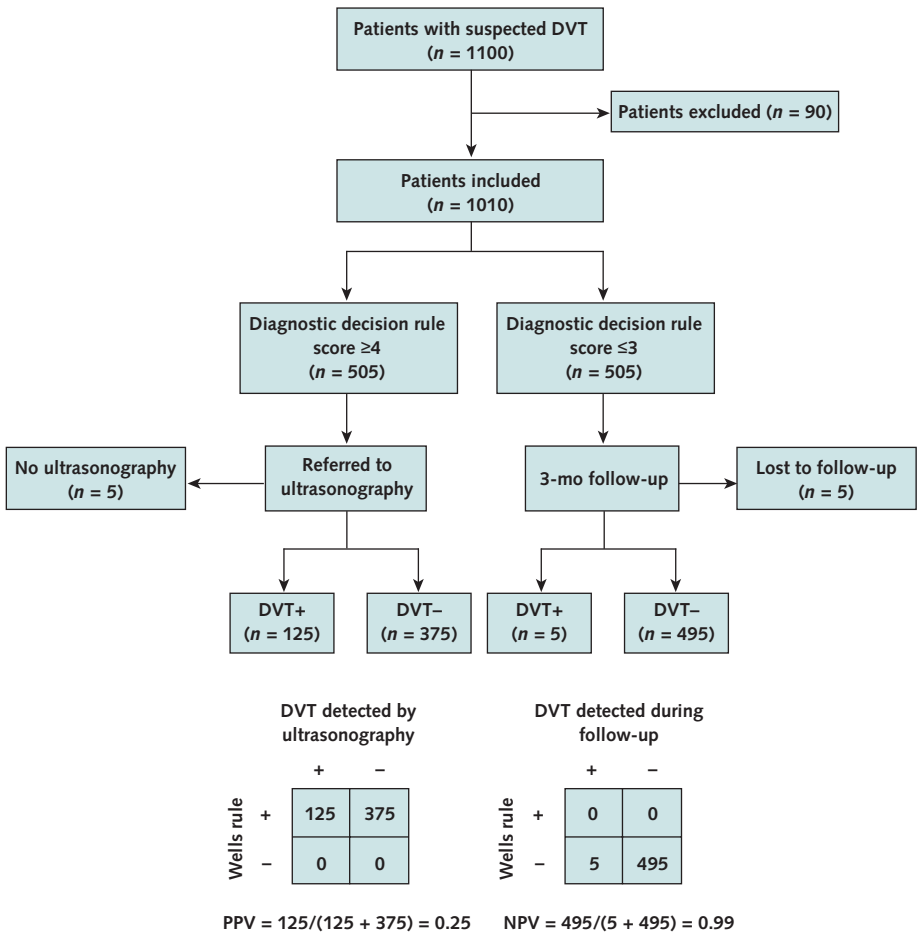
The pattern of verification plays an important role in whether bias is introduced. The most straightforward verification pattern is when the choice of reference standard is fully dependent on the index test results (**Figure 3**, *pattern A*). Studies with this pattern are likely to have biased estimates of sensitivity, specificity, diagnostic odds ratios, and likelihood ratios because these estimates rely partly on disease status classification by the inferior reference standard. An exception is positive predictive value estimates (the probability that a person has the disease given that the index test results are abnormal). The positive predictive value estimate is not affected by differential verification bias in pattern A because all patients with positive index test results receive the preferred reference standard (3). Negative predictive value estimates can also still be interpreted in a meaningful way in the sense that they provide information on the proportion of missed cases, as defined by the inferior reference standard (see **Figure 4** for an example).

When the choice of reference standard is not fully dependent on the index test results (**Figure 3**, *pattern B*), differential verification is likely to bias all accuracy estimates because they rely to some degree on disease classification by the inferior reference standard. In the rare case

Table 2. **Questions to Ask When Assessing Risk of Differential Verification Bias**

Was the choice of reference standard completely dependent on the results of the index test? (If so, the predictive values are clinically interpretable.)
If the answer to the first question is no, how accurate is the inferior reference standard? (The higher the accuracy of the inferior reference standard, the lower the risk of bias.)
What percentage of the participants were diagnosed by use of the inferior reference standard? (If a negligible percentage of participants received an inferior standard, the risk of bias is low. Several factors must be taken into account when determining whether the percentage is negligible.)
If follow-up is used as the inferior reference standard, does it identify almost all hidden cases present at the time of the index test but very few new cases that develop afterward? Does follow-up detect the same type of cases as the preferred reference standard? (If the answer to both questions is yes, the risk of bias is low.)

Figure 4. **An example of how to report complete index test–dependent differential verification.**



The example is inspired by AMUSE-1 (Amsterdam Maastricht Utrecht Study on Thromboembolism) (7), which looked at the safety of a diagnostic rule (the Wells rule) for excluding DVT. When the rule predicted a high risk for DVT, the patients were referred for further testing (ultrasonography). Those with a low predicted risk were sent home and followed instead. In this example, the probability that a participant truly has DVT given that they had a diagnostic score ≥4 (that is, the PPV) is 25%. Likewise, the probability that a person who was sent home because they had a diagnostic score ≤3 would soon thereafter be diagnosed with DVT (and was therefore erroneously sent home) was 1%. DVT = deep venous thrombosis; NPV = negative predictive value; PPV = positive predictive value.

that reference standard assignment is completely random, index test accuracy estimates calculated using only data from patients verified by the preferred reference standard will be unbiased, although they will generally lack precision because of the smaller sample size.

When the pattern of differential verification is not reported clearly, it is difficult for the reader to assess whether differential verification introduces clinically relevant bias. As a last resort, it is possible to take a pragmatic approach. This approach is to consider studies to have a low risk of bias if only a negligible percentage of participants receive the inferior reference standard. However, determining this percentage requires consideration of other factors that may influence this bias, particularly the degree of imperfection of the inferior reference standard and the correlation of errors between the inferior reference standard and the test under evaluation.

Because differential verification often biases diagnostic index test or model accuracy estimates, all efforts should be made to avoid it. When it is unavoidable, the fully index test–dependent differential verification pattern (**Figure 3**, *pattern A*) is preferable because it still produces *clinically interpretable* predictive values, as discussed earlier. Pattern B in **Figure 3** probably biases all index test accuracy estimates if analyzed naively.

Methods for Dealing With Differential Verification Bias

Adjusting for differential verification bias is possible when the indication to perform a particular reference standard can be explained by the data at hand. These methods can improve accuracy estimates, but the gains in accuracy are correlated with a loss in precision.

Verification pattern A in **Figure 3** is obvious to the investigator, can be confirmed by the data, and is probably well-explained to the reader. However, in pattern B of **Figure 3**, the indication to perform a particular reference standard may require assessment of the associations between observed variables in the data set and the choice of reference standard, similar to the way that the pattern of missingness is investigated in methods for handling missing data (17). The more the important patient characteristics differ between reference standard groups, the more likely differential verification introduces bias. However, if all key patient characteristics are found to be similar between the groups, it may, for the sake of practicality, be reasoned that reference standard assignment was a near-random process (17).

When the indication to perform a particular reference standard is known (or can be assumed to be completely random) and well-reported, the investigator or a reader can use a recently proposed Bayesian correction method (14). This model takes into account the verification pattern as well as bias due to *one or both* imperfect reference standards in a single model. The model requires the investigator to specify the verification pattern and give a best guess of the accuracy of *both* reference standards in the form of a prior distribution. The data (that is, the cross-tabulations of the index test results by the results of the reference standards) are also required.

When the preferred reference standard is not too burdensome for the patient, investigators might consider performing both reference standards in a random subset of participants (for pattern A in **Figure 3**, this could be done in a random subset of patients with negative index test results) to estimate the concordance, and thus the accuracy, of the inferior reference standard compared with the preferred reference standard. This information is helpful in the Bayesian method discussed earlier because it provides a better estimate of the accuracy of the imperfect reference standard (that is, a narrower prior distribution that is closer to the truth), which will result in more precise and less biased estimates. Another option is to use the information on the patients who received both reference standards to predict what the result of the preferred reference standard will be in those missing this result by using a method called multiple imputation (19).

Reporting Differential Verification in Diagnostic Studies

Identifying differential verification and assessing the associated risk of bias is often difficult because of insufficient reporting. In our view, a commonly used diagnostic accuracy reporting guideline, STARD, could be improved on in this aspect. Currently, this guideline recommends the inclusion of a flow diagram of the verification process as well as cross-tabulation of the index tests by reference standard (16). Examples of how to report results when differential verification is present could be included in future versions of STARD. These examples should include flow charts representing the differential verification process as well as a presentation of the results for each reference standard (see **Figure 4**).

When the index test is dichotomous, contingency tables comparing the index test with the reference standards should be reported for each reference standard to convey the pattern and extent of differential verification to the reader. For continuous diagnostic index tests or models, separate receiver-operating characteristic curves for the groups that were verified by the different reference standards should be presented. In addition to presenting study results in figures and tables, it is helpful to report reasons that groups of patients received the preferred or inferior reference standard as well as an estimate from the literature or an educated guess on the accuracy of the inferior reference standard. Detailed reporting of differential verification will help the reader to assess the risk of bias.

Assessing the Risk of Clinically Relevant Differential Verification Bias

The initial signaling question on differential verification bias in QUADAS-1 was, "Did the patients receive the same reference standard *regardless of the index test result*?" (25). This question proved difficult to answer because studies often do not report why patients received the preferred or the inferior reference standard. Consequently, this question was simplified in QUADAS-2 to, "Did all patients receive the same reference standard?" (13, 26). Although this question is easier to answer, it is unfortunately less probing. We propose the additional questions in **Table 2** for assessing the risk of clinically relevant differential verification bias.

Concluding Remarks

Differential verification refers to a common situation in diagnostic studies in which different reference standards are used for different groups of patients. The use of different methods to determine the final disease status often results in biased estimates of diagnostic test or model accuracy and should therefore be avoided, if possible. Because of insufficient reporting, readers may struggle with identifying differential verification and assessing the risk of clinically relevant bias. When differential verification occurs, we recommend that authors not only assess the risk of bias themselves and try to correct for it when appropriate but also allow the reader to do the same by following the suggested reporting guidelines. When readers assess whether differential verification caused clinically relevant bias in accuracy estimates, we recommend that they consider both the accuracy of the inferior reference standard and the verification pattern.

In situations where differential verification is unavoidable, it may be impossible to obtain sufficiently unbiased and precise accuracy estimates even after making adjustments in the analysis for the imperfectness of the inferior reference standard. This is the case when the reason that patients receive one reference standard over another is unclear (and also cannot be assumed to be completely at random) or when the accuracy of the inferior reference test is poor or highly uncertain. When traditional diagnostic accuracy studies are unable to produce precise and unbiased results, comparative accuracy studies or randomized trials that explore the relationship between testing and improved health outcomes may be preferable (27, 28). However, these studies also have limitations and are not always feasible. Our guidance for studies with differential verification can be used with STARD to improve the reporting of diagnostic accuracy studies and with QUADAS to improve the assessment of risk of bias due to differential verification in reviews of diagnostic accuracy studies.

From University Medical Center Utrecht, Utrecht, the Netherlands.

Financial Support: By the Netherlands Organization for Scientific Research (project 918.10.615).

Potential Conflicts of Interest: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M13-0322.

Requests for Single Reprints: Christiana A. Naaktgeboren, MPH, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, the Netherlands.

Current author addresses and author contributions are available at www.annals.org.

References

1. **Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM.** Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess. 2007;11:iii, ix-51. [PMID: 18021577]

2. **Knottnerus JA.** The Evidence Base of Clinical Diagnosis. London: BMJ; 2003.

3. **de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, et al.** Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ. 2011;343:d4770. [PMID: 21810869]

4. **Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al.** Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6. [PMID: 10493205]

5. **Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM.** Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174:469-76. [PMID: 16477057]

6. **Lehman CD, Lee CI, Loving VA, Portillo MS, Peacock S, DeMartini WB.** Accuracy and value of breast ultrasound for primary imaging evaluation of symptomatic women 30-39 years of age. AJR Am J Roentgenol. 2012;199:1169-77. [PMID: 23096195]

7. **Büller HR, Ten Cate-Hoek AJ, Hoes AW, Joore MA, Moons KG, Oudega R, et al; AMUSE (Amsterdam Maastricht Utrecht Study on thromboEmbolism) Investigators.** Safely ruling out deep venous thrombosis in primary care. Ann Intern Med. 2009;150:229-35. [PMID: 19221374]

8. **Thangaratinam S, Brown K, Zamora J, Khan KS, Ewer AK.** Pulse oximetry screening for critical congenital heart defects in asymptomatic newborn babies: a systematic review and meta-analysis. Lancet. 2012;379:2459-64. [PMID: 22554860]

9. **Zuithoff NP, Vergouwe Y, King M, Nazareth I, Hak E, Moons KG, et al.** A clinical prediction rule for detecting major depressive disorder in primary care: the PREDICT-NL study. Fam Pract. 2009;26:241-50. [PMID: 19546117]

10. **Gupta A, Chandrasekhar A, Gupte N, Patil S, Bhosale R, Sambarey P, et al; Byramjee Jeejeebhoy Medical College–Johns Hopkins University Study Group.** Symptom screening among HIV-infected pregnant women is acceptable and has high negative predictive value for active tuberculosis. Clin Infect Dis. 2011;53:1015-8. [PMID: 21940417]

11. **Alonzo TA, Brinton JT, Ringham BM, Glueck DH.** Bias in estimating accuracy of a binary screening test with differential disease verification. Stat Med. 2011;30:1852-64. [PMID: 21495059]

12. **Walter SD, Macaskill P, Lord SJ, Irwig L.** Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. Stat Med. 2012;31:1129-38. [PMID: 22351623]

13. **Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al; QUADAS-2 Group.** QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529-36. [PMID: 22007046]

14. **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. [PMID: 21228702]

15. **Gaffikin L, McGrath JA, Arbyn M, Blumenthal PD.** Visual inspection with acetic acid as a cervical cancer test: accuracy validated using latent class analysis. BMC Med Res Methodol. 2007;7:36. [PMID: 17663796]

16. **Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al; STARD Group.** Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Fam Pract. 2004;21:4-10. [PMID: 14760036]

17. **Little RJA, Rubin DB.** Statistical Analysis with Missing Data. New York: Wiley-Interscience; 2002.

18. **Begg CB, Greenes RA.** Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983;39:207-15. [PMID: 6871349]

19. **de Groot JA, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB.** Multiple imputation to correct for partial verification bias revisited. Stat Med. 2008;27:5880-9. [PMID: 18752256]

20. Differential Oxford Dictionaries Web site. Accessed at http://oxforddictionaries.com on 4 January 2013.

21. **Harel O, Zhou XH.** Multiple imputation for correcting verification bias. Stat Med. 2006;25:3769-86. [PMID: 16435337]

22. **Donders AR, van der Heijden GJ, Stijnen T, Moons KG.** Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087-91. [PMID: 16980149]

23. **Ewer AK, Furmston AT, Middleton LJ, Deeks JJ, Daniels JP, Pattison HM, et al.** Pulse oximetry as a screening test for congenital heart defects in newborn infants: a test accuracy study with evaluation of acceptability and cost-effectiveness. Health Technol Assess. 2012;16:v-xiii, 1-184. [PMID: 22284744]

24. **Lord SJ, Staub LP, Bossuyt PM, Irwig LM.** Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. BMJ. 2011;343:d4684. [PMID: 21903693]

25. **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. [PMID: 14606960]

26. **Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J.** Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9. [PMID: 16519814]

27. **Hayen A, Macaskill P, Irwig L, Bossuyt P.** Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. J Clin Epidemiol. 2010;63:883-91. [PMID: 20079607]

28. **Sonke GS, Verbeek AL, Kiemeney LA.** A philosophical perspective supports the need for patient-outcome studies in diagnostic test evaluation. J Clin Epidemiol. 2009;62:58-61. [PMID: 18619792]

Current Author Addresses: Ms. Naaktgeboren; Drs. de Groot, Moons, and Reitsma; and Mr. van Smeden: Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, the Netherlands.

Author Contributions: Conception and design: C.A. Naaktgeboren, J.A.H. de Groot, M. van Smeden, K.G.M. Moons, J.B. Reitsma.
Analysis and interpretation of the data: K.G.M. Moons, J.B. Reitsma.
Drafting of the article: C.A. Naaktgeboren, J.A.H. de Groot, K.G.M. Moons.
Critical revision of the article for important intellectual content: C.A. Naaktgeboren, J.A.H. de Groot, M. van Smeden, K.G.M. Moons, J.B. Reitsma.
Final approval of the article: J.A.H. de Groot, M. van Smeden, K.G.M. Moons, J.B. Reitsma.
Statistical expertise: J.A.H. de Groot, M. van Smeden, K.G.M. Moons, J.B. Reitsma.
Obtaining of funding: K.G.M. Moons.
Administrative, technical, or logistic support: J.A.H. de Groot.
Collection and assembly of data: K.G.M. Moons.