Adjusting for Bias in Diagnostic Reports

Majanka H. Heijenbrok-Kal, MSc, M. G. Myriam Hunink, MD, PhD

The performance of a diagnostic test can be evaluated by comparing its results with a reference standard. Measures of sensitivity, specificity, and accuracy can then be used as summary estimates of the test's performance. However, there are several limitations to the internal and external validity of a diagnostic test evaluation (1,2). Flaws in patient selection, study conduct, and data analysis can lead to biased results. Selection bias, referral bias, or spectrum bias may occur if patients are not included consecutively in the study; in which case, the selected patient group is not representative of the target group of the diagnostic test. There may be verification or work-up bias, which is a form of referral bias, if the test under evaluation influences the decision to perform the reference test, and if not all patients tested undergo the reference standard test. Disease progression may bias performance parameters if the time between the diagnostic test and the reference test is too long. Excluding indeterminate test results and patients lost to follow-up may lead to overestimation of the test characteristics. Intra- and interobserver variation, awareness of the results of the diagnostic test when interpreting the reference standard (or vice versa), and awareness of clinical characteristics may also influence test performance.

As long as reporting of the diagnostic test evaluation is complete, one can assess the potential for bias and the generalizability of the test results. The quality of reporting diagnostic test evaluations, however, is often suboptimal. Comparing publications reporting diagnostic test evaluations requires, at the very least, a complete description of the test, its positivity criterion, and the reference standard. Often, this information is lacking. Especially for the purpose of meta-analysis, reporting information on the design and conduct of diagnostic studies is essential, so that statistical adjustments can be made. Groups such as the Standards for Reporting of Diagnostic Accuracy Working Group have developed recommendations that may help accurate and complete reporting of diagnostic test evaluations (3).

In this issue of the *Journal*, Miller et al. deal with referral and verification bias concerning single photon emission computed tomography (SPECT) testing in patients with suspected coronary artery disease (4). Over 10 years, 14 273 patients from the Mayo Clinic underwent stress SPECT testing, of whom 13% were referred for coronary angiography. Twenty-six percent of patients with positive SPECT results underwent coronary angiography, whereas only 1% of the normal SPECT results were followed by angiography, suggesting that negative test results are less frequently verified by the reference standard test, and thus false-negative and true-negative results are likely to be missed, yielding high apparent sensitivity and low apparent specificity.

The authors, however, recognized the potential verification bias in their estimates and estimated test parameters adjusted for this bias. They used the correction method described by Begg and Greenes (1), as well as Diamond's correction method (5). The Begg and Greenes method adjusts for both pre- and post-test referral bias, thus correcting for both patient characteristics and the test result that together determine whether a patient will be referred for the reference test. The Diamond method adjusts only for post-test referral bias, that is, bias due to the influence that the test result has on the decision to refer the patient for the reference test. In addition, while the Begg and Greenes method can be applied to all types of receiver operating characteristic (ROC) curves, the Diamond method can only be applied to symmetrical ROC curves.

Miller et al. used a modified version of the Begg and Greenes method; they calculated the post-test probability or predictive value of coronary artery disease conditional on clinical factors and test results among patients who underwent coronary angiography. They assumed that the derived prediction equation would also apply to those who did not undergo coronary angiography. In essence, the assumption is that the predictive values of the diagnostic test are the same for the verified and source populations, which is equivalent to assuming that disease status and verification by the reference test are conditionally independent (1).

The effect of correcting for verification bias on the point estimates of test performance can best be shown using a meta-analysis of SPECT studies (6). Although one can never fully adjust for bias in the source studies, it is possible to evaluate several types of bias in meta-analyses. Presence of verification bias was not a significant predictor of test performance in the meta-analysis by Fleischmann et al. We superimposed the apparent and adjusted operating points that Miller et al. had reported on the summary ROC (SROC) curve from the meta-analysis (Figure). The SROC curve was adjusted for age to be representative of the patient population from Miller et al.'s study. Whereas the correction for verification bias has a

Am J Med. 2002;112:322–324.

From the Program for the Assessment of Radiological Technology, Department of Epidemiology and Biostatistics and the Department of Radiology, Erasmus Medical Center, Rotterdam, The Netherlands.

Requests for reprints should be addressed to M. G. Myriam Hunink, MD, PhD, Assessment of Radiological Technology Program, Department of Epidemiology and Biostatistics and the Department of Radiology, Erasmus Medical Center, Room EE21-40A, Dr Molewaterplein 50, 3015 GE Rotterdam, The Netherlands.

0002-9343/02/$–see front matter
PII S0002-9343(02)01042-2

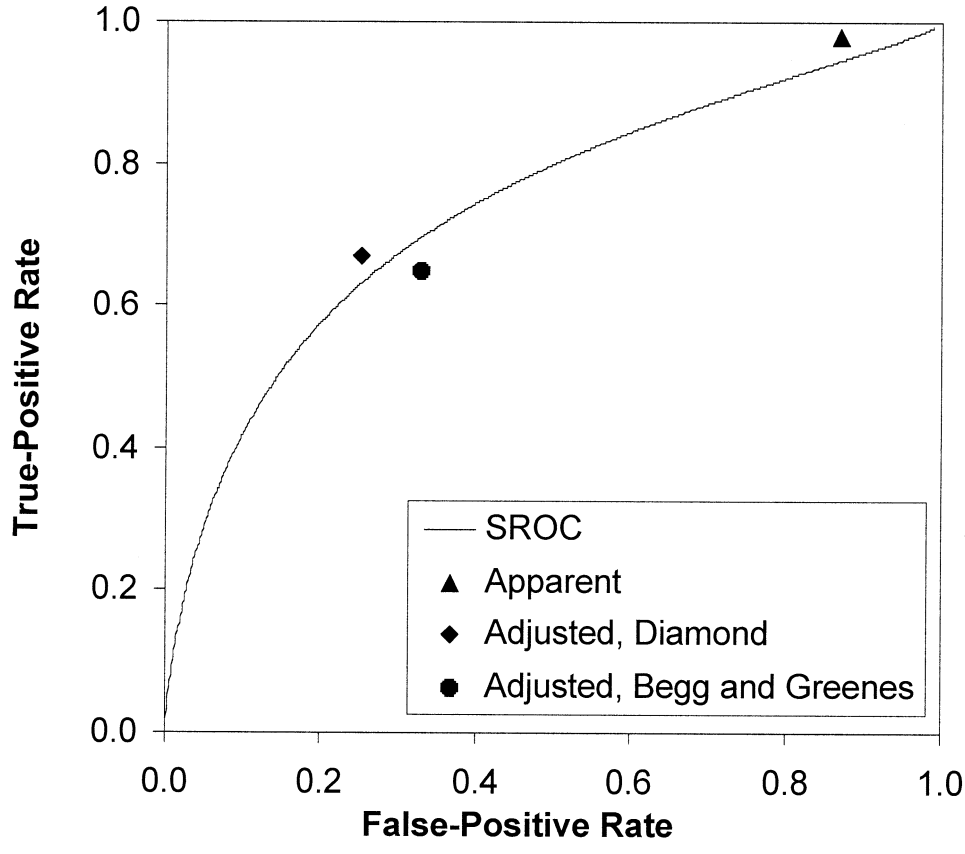

Figure. Apparent and adjusted point estimates (Diamond and Begg-Greenes methods) of the report by Miller et al. (4) superimposed on a summary receiver operating characteristic (SROC) curve of single photon emission computed tomography from a published meta-analysis (6).

large effect on the apparent point estimates of test performance, as reported by Miller et al., the apparent and corrected operating points lie along the same SROC curve (Figure), suggesting that the overall diagnostic performance as represented by an (S)ROC curve is not necessarily affected by verification bias, unlike the operating point that corresponds with a particular positivity criterion (threshold value). This, in turn, implies that when choosing a positivity criterion for a test, verification bias in the estimated test performance should be considered.

To determine the optimal positivity criterion (threshold value for calling a test score positive), we need to determine the optimal operating point on the (S)ROC curve. The optimal operating point on the (S)ROC curve represents the best combination of sensitivity and specificity for a patient population in terms of effectiveness and cost. The optimal operating point is determined by the prior probability of disease, the benefit of correctly diagnosing the disease compared with it remaining undetected (true positive compared with false negative), and the losses associated with incorrectly labeling a subject diseased compared with not doing so (false positive compared with true negative). In other words, we need to adjust our operating point so that it is consistent with the trade-offs in risks and benefits (7). Taking verification bias into account implies another adjustment when translating the derived optimal operating point to the threshold test score to be used in practice. After considering the risks and benefits, we may, for example, conclude that we want a true-positive rate of 0.80 with a corresponding false-positive rate of 0.50. Since the reported scoring system corresponded with an adjusted sensitivity of 0.67, we would need to use a more lenient criterion to achieve a sensitivity of 0.80. For example, instead of only considering reversible perfusion defects and moderate fixed defects as a positive test result, one could also consider mild fixed defects to diagnose coronary artery disease. In general, verification bias overestimates apparent sensitivity if positive test results are verified preferentially, which implies that to achieve a particular sensitivity we would need to use a more lenient criterion than it would seem based on the verified sample only. In diagnosing coronary artery disease, we could achieve this by adjusting the scoring system, or the threshold score, based on the SPECT images.

In summary, given that sensitivity and specificity esti-

mates depend on how the test is conducted, the setting, patient characteristics, and interpretation of test results, we should strive for accurately reporting how the test was evaluated and how the data were analyzed, so that bias can be recognized, validity can be judged, and appropriate adjustments can be made.

REFERENCES

1. Begg CB, Greenes A. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics.* 1983;39:207–215.
2. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999;282: 1061–1066.
3. Standards for Reporting of Diagnostic Accuracy (STARD) Working Group. The STARD Initiative-towards complete and accurate reporting of studies on diagnostic accuracy. Available at: http://www.consort-statement.org/stardstatement.htm.
4. Miller TD, Hodge DO, Christian TF, et al. Effects of adjustment for referral bias on the sensitivity and specificity of single photon emission computed tomography for the diagnosis of coronary artery disease. *Am J Med.* 2002;112:290–297.
5. Diamond GA, Rozanski A, Forrester JS, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. Application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. *J Chronic Dis.* 1986;39:343–355.
6. Fleischmann KE, Hunink MG, Kuntz KM, Douglas PS. Exercise echocardiography or exercise SPECT imaging? A meta-analysis of diagnostic test performance. *JAMA.* 1998;280:913–920.
7. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. *Med Decis Making.* 1988;8:279–289.