

Sources of Variation and Bias in Studies of Diagnostic Accuracy

A Systematic Review

Penny Whiting, MSc; Anne W.S. Rutjes, MSc; Johannes B. Reitsma, MD, PhD; Afina S. Glas, MD, PhD; Patrick M.M. Bossuyt, PhD; and Jos Kleijnen, MD, PhD

Background: Studies of diagnostic accuracy are subject to different sources of bias and variation than studies that evaluate the effectiveness of an intervention. Little is known about the effects of these sources of bias and variation.

Purpose: To summarize the evidence on factors that can lead to bias or variation in the results of diagnostic accuracy studies.

Data Sources: MEDLINE, EMBASE, and BIOSIS, and the methodologic databases of the Centre for Reviews and Dissemination and the Cochrane Collaboration. Methodologic experts in diagnostic tests were contacted.

Study Selection: Studies that investigated the effects of bias and variation on measures of test performance were eligible for inclusion, which was assessed by one reviewer and checked by a second reviewer. Discrepancies were resolved through discussion.

Data Extraction: Data extraction was conducted by one reviewer and checked by a second reviewer.

Data Synthesis: The best-documented effects of bias and variation were found for demographic features, disease prevalence and severity, partial verification bias, clinical review bias, and observer and instrument variation. For other sources, such as distorted selection of participants, absent or inappropriate reference standard, differential verification bias, and review bias, the amount of evidence was limited. Evidence was lacking for other features, including incorporation bias, treatment paradox, arbitrary choice of threshold value, and dropouts.

Conclusions: Many issues in the design and conduct of diagnostic accuracy studies can lead to bias or variation; however, the empirical evidence about the size and effect of these issues is limited.

Ann Intern Med. 2004;140:189-202.

For author affiliations, see end of text.

www.annals.org

Diagnostic tests are of crucial importance in health care. They are performed to reduce uncertainty concerning whether a patient has a condition of interest. A thorough evaluation of diagnostic tests is necessary to ensure that only accurate tests are used in practice. Diagnostic accuracy studies are a vital step in this evaluation process.

Diagnostic accuracy studies aim to investigate how well the results from a test being evaluated (index test) agree with the results of the reference standard. The reference standard is considered the best available method to establish the presence or absence of a condition (target condition). In a classic diagnostic accuracy study, a consecutive series of patients who are suspected of having the target condition undergo the index test; then, all patients are verified by the same reference standard. The index test and reference standard are then read by persons blinded to the results of each, and various measures of agreement are calculated (for example, sensitivity, specificity, likelihood ratios, and diagnostic odds ratios).

This classic design has many variations, including differences in the way patients are selected for the study, in test protocol, in the verification of patients, and in the way the index test and reference standard are read. Some of these differences may bias the results of a study, whereas others may limit the applicability of results. Bias is said to be present in a study if distortion is introduced as a consequence of defects in the design or conduct of a study. Therefore, a biased diagnostic accuracy study will produce estimates of test performance that differ from the true performance of the test. In contrast, variability arises from differences among studies, for example, in terms of popu-

lation, setting, test protocol, or definition of the target disorder (1). Although variability does not lead to biased estimates of test performance, it may limit the applicability of results and thus is an important consideration when evaluating studies of diagnostic accuracy.

The distinction between bias and variation is not always straightforward, and the use of different definitions in the literature further complicates this issue. For example, when a diagnostic study starts by including patients who have already received a diagnosis of the target condition and uses a group of healthy volunteers as the control group, it is likely that both sensitivity and specificity will be higher than they would be in a study made up of patients only suspected of having the target condition. This feature has been described as spectrum bias. However, strictly speaking, one could argue that it is a form of variability; sensitivity and specificity have been measured correctly within the study and thus there is no bias; however, the results cannot be applied to the clinical setting. In other words, they lack generalizability (2). Others have argued that when the goal of a study is to measure the accuracy of a test in the clinical setting, an error in the method of patient selection is made that will lead to biased estimates of test performance. They use a broader definition of bias and take into account the underlying research question when deciding whether results are biased. In this paper, we use a more restricted definition of bias.

Our goal is to classify the various sources of variation and bias, describe their effects on test results, and provide a summary of the available evidence that supports each source of bias and variation (Table 1). For this purpose, we

Table 1. Description of Sources of Bias and Variation

Source	Bias or Variation	Description
Population		
Demographic features	Variation	Tests may perform differently in various samples. Therefore, demographic features may lead to variations in estimates of test performance.
Disease severity	Variation	Differences in disease severity among studies may lead to differences in estimates of test performance.
Disease prevalence	Variation	The prevalence of the target condition varies according to setting and may affect estimates of test performance. Context bias, the tendency of interpreters to consider test results to be positive more frequently in settings with higher disease prevalence, may also affect estimates of test performance.
Distorted selection of participants	Variation	The selection process determines the composition of the study sample. If the selection process does not aim to include a patient spectrum similar to the population in which the test will be used in practice, the results of the study may have limited applicability.
Test protocol: materials and methods		
Test execution	Variation	A sufficient description of the execution of index and reference standards is important because variation in measures of diagnostic accuracy can be the result of differences in test execution.
Test technology	Variation	When the characteristics of a diagnostic test change over time as a result of technological improvement or the experience of the operator of the test, estimates of test performance may be affected.
Treatment paradox and disease progression bias	Bias	Disease progression bias occurs when the index test is performed an unusually long time before the reference standard, so the disease is at a more advanced stage when the reference standard is performed. Treatment paradox occurs when treatment is started on the basis of the knowledge of the results of the index test, and the reference standard is applied after treatment has started.
Reference standard and verification procedure		
Inappropriate reference standard	Bias	Errors of imperfect reference standard or standards bias the measurement of diagnostic accuracy of the index test.
Differential verification bias	Bias	Part of the index test results is verified by a different reference standard.
Partial verification bias	Bias	Only a selected sample of patients who underwent the index test is verified by the reference standard.
Interpretation (reading process)		
Review bias	Bias	Interpretation of the index test or reference standard is influenced by knowledge of the results of the other test. Diagnostic review bias occurs when the results of the index test are known when the reference standard is interpreted. Test review bias occurs when results of the reference standard are known while the index test is interpreted.
Clinical review bias	Bias	The availability of information on clinical data, such as age, sex, and symptoms, during interpretation of test results may affect estimates of test performance.
Incorporation bias	Bias	The result of the index test is used to establish the final diagnosis.
Observer variability	Variation	The reproducibility of test results is one of the determinants of diagnostic accuracy of an index test. Because of variation in laboratory procedures or observers, a test may not consistently yield the same result when repeated. In 2 or more observations of the same diagnostic study, intraobserver variability occurs when the same person obtains different results, and interobserver variability occurs when 2 or more people disagree.
Analysis		
Handling of indeterminate results	Bias	A diagnostic test can produce an uninterpretable result with varying frequency depending on the test. These problems are often not reported in test efficacy studies; the uninterpretable results are simply removed from the analysis. This may lead to biased assessment of the test characteristics.
Arbitrary choice of threshold value	Variation	The selection of the threshold value for the index test that maximizes the sensitivity and specificity of the test may lead to overoptimistic measures of test performance. The performance of this cutoff in an independent set of patients may not be the same as in the original study.

conducted a systematic review of all studies in which the main focus was examine the effects of one or more sources of bias or variation on estimates of test performance.

METHODS

Literature Searches

We searched MEDLINE, EMBASE, BIOSIS and the methodologic databases of the Centre for Reviews and Dissemination and the Cochrane Collaboration from database inception to 2001. Search terms included *sensitivity**, *mass-screening*, *diagnostic-test*, *laboratory-diagnosis*, *false positive**, *false negative**, *specificity**, *screening*, *accuracy*, *predictive value**, *reference value**, *likelihood ratio*, *sroc*, and *receiver op-*

erat characteristic**. We also identified papers that had cited the key papers. Complete details of the search strategy are provided elsewhere (3). We contacted methodologic experts and groups conducting work in this field. Reference lists of retrieved articles were screened for additional studies.

Inclusion Criteria

All studies with the main objective of addressing bias or variation in the results of diagnostic accuracy studies were eligible for inclusion. Studies of any design, including reviews, and any topic area were eligible. Studies had to investigate the effects of bias or variation on measures of test performance, such as sensitivity, specificity, predictive

values, likelihood ratios, and diagnostic odds ratios, and indicate how a particular feature may distort these measures. Inclusion was assessed by one reviewer and checked by a second reviewer; discrepancies were resolved through discussion.

Data Extraction

One reviewer extracted data and a second reviewer checked data on the following parameters: study design, objective, sources of bias or variation investigated, and the results for each source. Discrepancies were resolved by consensus or consultation with a third reviewer.

Data Synthesis

We divided the different sources of bias and variation into groups (Table 1). Table 1 provides a brief description of each source of bias and variation; more detailed descriptions are available elsewhere (3). Results were stratified according to the source of bias or variation. Studies were grouped according to study design. We classified studies that used actual data from one or more clinical studies to demonstrate the effect of a particular study feature as experimental studies, diagnostic accuracy studies, or systematic reviews. Experimental studies were defined as studies specifically designed to test a hypothesis about the effect of a certain feature, for example, rereading sets of radiographs while controlling (manipulating) the overall prevalence of abnormalities. Studies that used models to simulate how certain types of biases may affect estimates of diagnostic test performance were classified as modeling studies. These studies were considered to provide theoretical evidence of bias or variation.

Role of the Funding Source

The funding source was not involved in the design, conduct, or reporting of the study or in the decision to submit the manuscript for publication.

DATA SYNTHESIS

The literature searches identified a total of 8663 references. Of these, 569 studies were considered potentially relevant and were assessed for inclusion; 55, published from 1963 to 2000, met inclusion criteria. Nine studies were systematic reviews, 16 studies used an experimental design, 22 studies were diagnostic accuracy studies, and 8 studies used modeling to investigate the theoretical effects of bias or variation.

Population

Demographic Features

Ten studies assessed the effects of demographic features on test performance (Table 2) (4, 5, 7, 9, 11, 14, 15, 20, 22, 24). Eight studies were diagnostic accuracy studies, and 2 were systematic reviews. All but one study (22) found an association between the features investigated and overall accuracy. The study that did not find an association investigated whether estimates of exercise testing performance differed between men and women; after correction

for the effects of verification bias, no significant differences were found (22).

In general, the studies found associations between the demographic factors investigated and sensitivity; the reported effect on specificity was less strong. Four studies found that various factors, including sex, were associated with sensitivity but showed no association with specificity (4, 5, 11, 20). The index tests investigated in these studies were exercise testing (5, 11, 20) to diagnose heart disease and body mass index to test for obesity (4). Two additional studies of exercise testing also reported an association with sensitivity, but the effects on specificity differed. One found that factors that lead to increased sensitivity also lead to a decrease in specificity (14); the second reported higher sensitivity and specificity in men than in women (16). A study of the diagnostic accuracy of an alcohol screening questionnaire found that overall accuracy was increased in certain ethnic groups (24). Sex was the most commonly investigated variable. Three studies found no association between test performance and sex, 9 found significant effects on sensitivity, and 4 found significant effects on specificity. Other variables shown to have significant effects on test performance were age, race, and smoking status.

Disease Severity

Six studies looked at the effects of disease severity on test performance (Table 2) (5, 11, 14, 19, 23, 25). Three studies were diagnostic accuracy studies, 2 were reviews, and one used modeling to investigate the effects of differences in disease severity. The modeling study also included an example from a diagnostic accuracy study of tests for the diagnosis of ovarian cancer (25). Three studies investigated tests for heart disease (5, 11, 14), one examined ventilation-perfusion lung scans for diagnosing pulmonary embolism (23), and one investigated 2 different laboratory tests (one for cancer and the other for bacterial infections) (19). All 6 studies found increased sensitivity with more severe disease; 5 found no effect on specificity (5, 11, 14, 19, 23), and one did not comment on the effects on specificity (25).

Disease Prevalence

Six studies looked at the effects of increased disease prevalence on test performance (Table 2) (8, 10, 13, 17, 21, 26). One study used an experimental design (8); the other studies were all diagnostic accuracy studies. The tests investigated in these studies covered a wide range of topics: dipstick for diagnosing urinary tract infection (10), magnetic resonance imaging and evoked potentials for diagnosing multiple sclerosis (17), exercise testing for diagnosing coronary artery disease (21), lung scans for diagnosing pulmonary embolism (8), clinical indications for diagnosing pneumonia (13), and ultrasonography for diagnosing epididymitis (26). Only 5 of the studies reported on the effects of disease prevalence on sensitivity; all found an in-

Table 2. Population*

Study, Year (Reference)	Design	Index Test	Study Sample
Curtin et al., 1997 (4)	Diagnostic accuracy	Body mass index	226 white persons
Detrano et al., 1988 (5)	Review	Exercise thallium scintigraphy	56 primary studies
Detrano et al., 1988 (6)			
Detrano et al., 1989 (7)	Review	Exercise electrocardiography	60 primary studies
Egglin and Feinstein, 1996 (8)	Experimental	Pulmonary arteriography	24 arteriograms
Hlatky et al., 1984 (9)	Diagnostic accuracy	Exercise electrocardiography	2269 patients
Lachs et al., 1992 (10)	Diagnostic accuracy	Dipsticks	366 consecutive patients
Levy et al., 1990 (11)	Diagnostic accuracy	Electrocardiography	4684 patients with suspected left ventricular hypertrophy
Lijmer et al., 1999 (12)	Review	Various tests	184 primary studies of 218 tests
Melbye and Straume, 1993 (13)	Diagnostic accuracy	Clinical cues	581 patients with suspected pneumonia
Moons et al., 1997 (14)	Diagnostic accuracy	Exercise test	295 consecutive patients with heart pain
Morise and Diamond, 1994 and 1995 (15, 16)	Diagnostic accuracy	Exercise electrocardiography	4467 patients with suspected coronary disease
O'Connor et al., 1996 (17)	Diagnostic accuracy	Magnetic resonance imaging and evoked potentials	303 patients with suspected multiple sclerosis
Philbrick et al., 1982 (18)	Diagnostic accuracy	Graded exercise test	208 consecutive patients evaluated for coronary arterial disease
Ransohoff and Feinstein, 1978 (19)	Review	Carcinoembryonic antigen and nitroblue tetrazolium tests	17 studies of carcinoembryonic antigen and 16 of nitroblue tetrazolium
Roger et al., 1997 (20)	Diagnostic accuracy	Exercise echocardiography	3679 consecutive patients
Rozanski et al., 1983 (21)	Diagnostic accuracy	Exercise radionuclide ventriculography	77 angiographically normal patients
Santana-Boado et al., 1998 (22)	Diagnostic accuracy	Single-photon emission computed tomography	702 consecutive patients evaluated for coronary disease
Stein et al., 1993 (23)	Diagnostic accuracy	Ventilation/perfusion scan	1050 patients
Steinbauer et al., 1998 (24)	Diagnostic accuracy	Screening tests for alcohol abuse	1333 adult family practice patients
Taube and Tholander, 1990 (25)	Modeling and diagnostic accuracy	Tests for epithelial ovarian cancer	168 patients with ovarian carcinoma
van der Schouw et al., 1995 (26)	Diagnostic accuracy	Ultrasonography	483 consecutive patients; 372 included
Van Rijkom et al., 1995 (27)	Review	Tests for approximal caries	39 sets of sensitivity and specificity data

* NA = not applicable; ↑ = increased; ↓ = decreased.

crease in sensitivity with increased disease prevalence (8, 10, 13, 17, 26). These studies also investigated the effects of increased disease prevalence on specificity and found mixed results; 2 found that specificity decreased (10, 13), 2 found no effect (8, 17), and one reported increased specificity (26). The remaining study looked only at the effects of disease prevalence on specificity, which was found to decrease (21).

Distorted Selection of Participants

Four studies examined the effects of distorted selection of participants on test performance (Table 2) (5, 12, 18,

27). A diagnostic accuracy study of exercise testing for heart disease found that overall accuracy was overestimated if reasons for exclusion commonly used by researchers were applied (18). The other 3 studies were reviews. The first, a review of the clinical and radiologic diagnosis of caries, found that in vivo studies gave higher estimates of test performance than in vitro studies (27). A review of exercise testing for heart disease found that avoiding a limited challenge group (that is, including patients with other confounding diseases or patients taking medications thought to produce false-positive results) did not have significant

Table 2—Continued

Source of Bias or Variation	Factors Investigated	Effect on Sensitivity	Effect on Specificity	Effect on Overall Accuracy
Demographic features	Increased weight; sex (female)	↑	None	NA
Demographic features	Sex, age, and medication use	Associated	None	NA
Distorted selection of participants	Avoidance of limited challenge group	None	None	NA
Disease severity	Inclusion of patients with previous myocardial infarction	↑	None	NA
Demographic features	Various patient-related characteristics (all are not associated)	Associated	Associated	NA
Disease prevalence	Context of interpretation: effect of increased disease prevalence	↑	None	NA
Demographic features	Exercise heart rate, number of diseased arteries, type of angina, age, and sex	Associated	Associated	NA
Disease prevalence	High pretest probability of disease	↑	↓	NA
Demographic features	Sex (male), increased age, decreased body mass index, not smoking	↑	None	NA
Disease severity	Increased severity of left ventricular hypertrophy	↑	None	NA
Distorted selection of participants	Diagnostic case-control studies	NA	NA	↑
	Nonconsecutive patient enrollment	NA	NA	None
	Retrospective study design	NA	NA	None
	Failure to describe patient spectrum	NA	NA	↑
Disease prevalence	Increased prevalence	↑	↓	NA
Demographic features	Sex, workload, diabetes, smoking, cholesterol level (all are not associated)	↑	↓	NA
Disease severity	Number of diseased vessels	↑	None	NA
Demographic features	Men	↑	↑	NA
Disease prevalence	Increased prevalence	↑	None	NA
Distorted selection of participants	Exclusion of patients with other clinical conditions	NA	NA	↑
Disease severity	Extensive disease	↑	None	NA
Demographic features	Sex (male)	↑	None	NA
Disease prevalence	Increased prevalence	Not reported	↓	NA
Demographic features	Sex	None	None	NA
Disease severity	Previous pulmonary disease	↑	None	NA
Demographic features	Race and sex	NA	NA	Associated
Disease severity	Clear cases of malignant disease	↑	Not reported	NA
Disease prevalence	Increased prevalence (inclusion criteria widened)	↑	↑	NA
Distorted selection of participants	In vivo studies compared with in vitro studies	NA	NA	↑

effects on overall accuracy (5). The final study, which reviewed many different tests, found that case-control studies overestimate overall accuracy; it also found that nonconsecutive patient enrollment and a retrospective study design did not affect the diagnostic odds ratio (12). This review also looked at the effects of failure to provide an appropriate description of the patient sample and found that this was associated with increased overall accuracy.

Test Protocol: Materials and Methods

Test Execution

We found only 2 studies, both reviews, that specifically looked at the effects of differences in test execution

(Table 3) (6, 12). The first, a review of several different tests, found that failure to describe the index test and reference standard execution leads to an overestimation of overall accuracy (12). The other found no effect of differences in protocol on overall accuracy in exercise testing (6).

Test Technology

Two studies looked at the effects of a change in the technology of the index test on test performance (Table 3) (6, 28). A systematic review of exercise scintigraphy studies found that automation of the test procedure improved sensitivity but decreased specificity (6). The other study, a

Table 3. Test Protocol: Materials and Methods*

Study Details	Design	Index Test	Study Sample	Source of Bias or Variation	Factors Investigated	Effect on Sensitivity	Effect on Specificity	Effect on Overall Accuracy
Detrano et al., 1988 (6)	Review	Exercise electrocardiography	60 primary studies	Test execution Test technology Disease progression bias	Exercise protocol Automation of test Maximum interval between scintigraphy and angiography	None ↑ None	None ↓ None	NA NA NA
Froelicher et al., 1998 (28)	Diagnostic accuracy	Electrocardiography and angiographic calipers	814 consecutive patients with angina pectoris	Test technology	Computerized readings	None	None	NA
Lijmer et al., 1999 (12)	Review	Various tests	184 primary studies of 218 tests	Test execution	Failure to describe index test execution; failure to describe reference standard execution	NA NA	NA NA	↑ ↓

* NA = not applicable; ↑ = increased; ↓ = decreased.

diagnostic accuracy study of the electrocardiographic exercise test, found no effect on test performance (28).

Treatment Paradox and Disease Progression Bias

No studies that provided evidence of the effect of treatment paradox were identified. Only one study that looked at the effects of disease progression bias on test performance was found. This study, a review of exercise scintigraphy for the diagnosis of heart disease, found no evidence of bias (6).

Reference Standard and Verification Procedure Inappropriate Reference Standard

Eight studies looked at reference standard error bias (Table 4) (6, 7, 27, 29, 31, 34, 41, 43). Four were systematic reviews, and the other 4 used modeling to investigate the theoretical effects of an imperfect reference standard. The reviews looked at reference standard error bias from slightly different perspectives, but all found evidence of bias. A review of patients who received a diagnosis of caries found that weaker validation methods may overestimate overall accuracy (27). A review of a hormone test for the diagnosis of depression found that different reference standards can provide very different estimates of sensitivity (29). A review of exercise scintigraphy for the diagnosis of heart disease found that studies that used a specific reference standard (tomographic imaging) overestimated sensitivity and specificity compared with other studies (6). The last review, which dealt with exercise electrocardiography for heart disease, found that comparison with a more accurate test leads to increased sensitivity but did not report on the effect on specificity (7).

The studies that used modeling to investigate the effects of an imperfect reference standard also found evidence of bias. One study suggested that with imperfect reference standards, specificity is most accurately estimated at low disease prevalence and sensitivity at high disease prevalence; it also suggested that considerable errors in estimates exist, even when the reference standard has close to perfect performance (31). Two studies found that inaccurate

reference standards lead to underestimation of index test accuracy when the index test errors are statistically independent of the reference standard and overestimation when the index test errors are statistically dependent on the reference standard (41, 43). The final study found that overall accuracy is underestimated when the test being evaluated is more accurate than the reference standard (34, 43).

Differential Verification Bias

Only 2 studies looked at differential verification bias (Table 4) (12, 30). One was a review of several different tests (12), and the other was a diagnostic accuracy study of the clinical diagnosis of Alzheimer disease (30). Both found that differential verification bias leads to higher (inflated) measures of overall accuracy.

Partial Verification Bias

Twenty studies investigated the effects of partial verification bias (Table 4) (5, 7, 12, 16, 18–22, 28, 30, 32, 35–40, 42, 44). Two studies used models to investigate the theoretical effects of verification bias and found that partial verification bias increased sensitivity and decreased specificity (35, 36). A third study also used modeling to investigate the effects of verification bias; in addition, it provided an example from a diagnostic accuracy study. This study reported an association between overall accuracy and the presence of partial verification bias (44).

All of the remaining studies used actual data to investigate the effects of partial verification bias and were either diagnostic accuracy studies or reviews. Most of these studies examined some form of exercise testing for the diagnosis of heart disease (5, 6, 16, 18, 20, 21, 28, 32, 38). Other tests that were investigated included noninvasive tests for arterial disease (37), clinical diagnosis for Alzheimer disease (30), clinical findings for diagnosing hemorrhage in patients who had strokes (40), nuchal translucency for diagnosing Down syndrome (39), the carcinoembryonic antigen and nitro-blue tests (19), and serum ferritin levels for diagnosing hereditary hemochromatosis (42). Seven studies

found that sensitivity was increased and specificity decreased in the presence of partial verification bias (16, 18, 20, 28, 32, 38, 40); one study found that both sensitivity and specificity were increased (39), and 2 studies found that sensitivity was increased but did not report the effects on specificity (19, 42). One study found that specificity was increased in the presence of verification bias (5) and another study reported that verification bias decreased specificity (21). Neither of these studies reported on the effects on sensitivity. Two studies did not report on the effects of partial verification bias on sensitivity and specificity. One of these found that partial verification bias increased overall accuracy (37), and the second reported that there was “scope for verification bias” but provided no additional information (30).

Two more studies found no evidence of bias. One was a systematic review of studies of the diagnostic accuracy of exercise electrocardiography (45), and the other was a review of systematic reviews of several different tests (12). The latter study used the relative diagnostic odds ratio as the summary statistic. If partial verification bias tends to increase sensitivity and decrease specificity, as is suggested by some of the studies, then no effect on the diagnostic odds ratio would be expected. This may explain why this review did not find any evidence of partial verification bias.

Interpretation (Reading Process)

Review Bias

Four studies investigated review bias (6, 12, 19, 45), 3 (6, 19, 45) examined diagnostic and test review bias, and one looked only at diagnostic review bias (Table 5) (12). A review of exercise testing found no effect of either diagnostic or test review bias on sensitivity and specificity (7). A separate review of exercise testing reported that both diagnostic and test review bias led to an increase in sensitivity but had no effect on specificity (5). A study of carcinoembryonic antigen and nitro-blue tests found that failure to avoid review bias may overestimate sensitivity and specificity (19). A review of several different tests looked only at diagnostic review bias and found that it increased overall accuracy (12).

Clinical Review Bias

Nine studies looked at the effects of clinical review bias (Table 5) (28, 46, 52, 53, 55–57, 59, 61). Most of these studies examined radiography (46, 52, 56, 57, 61), mammography (55), and myelography and spinal computer tomography (53). Eight studies used an experimental design, and one was a diagnostic accuracy study (28). One found no difference in overall accuracy between tests interpreted with and without clinical history (56). The other studies all found evidence of bias; however, the direction of bias differed among studies. In general, studies found that providing clinical information improved overall accuracy. Six studies reported that sensitivity was increased when clinical information was available (28, 46, 52, 53, 57, 61).

The effects of providing clinical information on specificity varied among these studies: Two reported that specificity decreased (52, 53), 2 found no effect on specificity (46, 61), and the other 2 did not report on the effects on specificity (28, 57). The remaining 2 studies did not report on the effects of providing clinical history on sensitivity and specificity, but both found that overall accuracy was improved when clinical information was provided (55, 59).

Incorporation Bias

No studies that looked at the effects of incorporation bias were identified.

Observer Variability

Eight studies looked at observer variation; no studies addressed instrument variation (Table 5) (47–51, 54, 58, 60). All studies used an experimental design. Most studies were evaluations of imaging techniques: radiologic detection of fractures (47), mammography (48, 54), and myocardial imaging (51). Other techniques that were evaluated were fine-needle aspiration biopsy (49), histologic examination (50), cytologic examination (60), and bronchial brush specimens (58). All 8 studies found evidence of interobserver variability, and 2 found evidence of intra-observer variability (48, 50); one of these studies reported that interobserver variability was greater than intraobserver variability (48). Two studies found that more experienced reviewers, or experts, provided greater sensitivity (49, 60), whereas another found that experience was not related to interobserver variability (58).

Analysis

Handling of Indeterminate Results

Two studies looked at the effects of uninterpretable test results (Table 6) (7, 18). One of these studies stated that a large proportion of results would be excluded if unsatisfactory test results were excluded but provided no evidence on how this may lead to biased estimates of test performance (18). The other study found that the treatment of equivocal or nondiagnostic test results was not associated with overall accuracy (7).

Arbitrary Choice of Threshold Value

No studies that provided evidence of the effect of the choice of threshold value were identified.

DISCUSSION

The searches identified a relatively small number of studies that looked specifically at the effects of bias and variation on estimates of diagnostic test performance. These studies were concentrated in 7 areas of bias and variation: demographic features (10 studies), disease prevalence (6 studies), disease severity (6 studies), inappropriate reference standard (8 studies), partial verification bias (20 studies), clinical review bias (9 studies), and observer variation (8 studies). Other sources of bias commonly believed

Table 4. Reference Standard and Verification Procedure*

Study Details	Design	Index Test	Study Sample
Arana et al., 1990 (29)	Review	Thyrotropin-releasing hormone stimulation	10 studies
Bowler et al., 1998 (30)	Diagnostic accuracy	Necropsy	307 patients
Boyko et al., 1988 (31)	Modeling	NA	Formulas used to model theoretical effects
Cecil et al., 1996 (32)	Diagnostic accuracy	Stress single-photon emission computed tomography thallium testing	4354 records selected from computerized database
De Neef, 1987 (34)	Modeling	New rapid antigen detection tests	Models used to vary reference standard accuracy
Detrano et al., 1988 (5, 6)	Review	Exercise thallium scintigraphy	56 primary studies
Detrano et al., 1989 (7)	Review	Exercise electrocardiography	60 primary studies
Diamond, 1991 (35)	Modeling	NA	Series of computer simulations using the Begg-Greenes method†
Diamond, 1992 (36)	Modeling	NA	Series of computer simulations using Bayes theorem
Froelicher et al., 1998 (28)	Diagnostic accuracy	Electrocardiography and angiographic calipers	814 consecutive patients with angina
Lijmer et al., 1999 (12)	Review	Various tests	184 primary studies of 218 tests
Lijmer et al., 1996 (37)	Diagnostic accuracy	Noninvasive tests	464 consecutive patients with suspected disease
Miller et al., 1998 (38)	Diagnostic accuracy	Stress imaging	15 945 low-risk patients
Mol et al., 1999 (39)	Review	Nuchal translucency measurement	25 studies
Morise and Diamond, 1994 and 1995 (15, 16)	Diagnostic accuracy	Exercise electrocardiography	4467 patients with suspected coronary disease
Panzer et al., 1987 (40)	Diagnostic accuracy	Clinical findings	374 patients with stroke and focal deficits
Phelps and Hutson, 1995 (41)	Modeling	NA	Monte Carlo studies
Philbrick et al., 1982 (18)	Diagnostic accuracy	Graded exercise test	208 consecutive patients
Ransohoff and Muir, 1982 (42)	Review	Serum ferritin levels	2 studies
Ransohoff et al., 1978 (19)	Review	Carcinoembryonic antigen and nitroblue tetrazolium tests	17 studies of carcinoembryonic antigen and 16 of nitroblue tetrazolium
Roger et al., 1997 (20)	Diagnostic accuracy	Exercise echocardiography	3679 consecutive patients
Rozanski et al., 1983 (21)	Diagnostic accuracy	Exercise ventriculography	77 angiographically normal patients
Santana-Boado et al., 1998 (22)	Diagnostic accuracy	Single-photon emission computed tomography	702 consecutive low-risk patients
Thibodeau, 1981 (43)	Modeling	NA	Various statistical models
van Rijkom and Verdonshot, 1995 (27)	Review	Tests for approximal caries	39 sets of sensitivity and specificity data
Zhou, 1994 (44)	Modeling and diagnostic accuracy	NA	429 patients

* DSM-III = *Diagnostic and Statistical Manual of the American Psychological Association*, 3rd edition; NA = not applicable; RDC = Research Diagnostic Criteria; ↑ = increased; ↓ = decreased.

† From Begg C and Greenes R (33).

to affect studies of diagnostic test performance, such as incorporation bias, treatment paradox, arbitrary choice of threshold value, and dropouts, were not considered in any studies.

Population

The evidence shows that differences in populations affect estimates of diagnostic performance. However, the extent and direction of the effect of variations in a population can vary, even among studies of the same index test.

Demographic features have shown strong associations with test performance and generally showed a greater effect on estimates of sensitivity than on specificity. Studies that observed effects on specificity generally found that factors that increased sensitivity also decreased estimates of specificity. There was also evidence that both disease severity and prevalence may affect estimates of test performance. Sensitivity tended to be increased in populations with more

Table 5. Interpretation (Reading Process)*

Study Details	Design	Index Test	Study Sample
Arana et al., 1990 (29)	Review	Thyrotropin-releasing hormone stimulation	10 studies
Berbaum et al., 1988 (46)	Experimental	Radiography	40 radiographs examined with and without clinical information
Berbaum et al., 1989 (47)	Experimental	Radiography	40 radiographs examined by a group of radiologists and a group of orthopedic surgeons
Ciccone et al., 1992 (48)	Experimental	Mammography	45 mammograms; 7 radiologists
Cohen et al., 1987 (49)	Experimental	Fine-needle aspiration biopsy	50 specimens examined by 5 observers
Corley et al., 1997 (50)	Experimental	Histologic diagnosis of pneumonia	39 lung biopsy samples, 4 pathologists
Cuaron et al., 1980 (51)	Experimental	Tc 99m phosphate myocardial imaging	250 myocardial slides evaluated by 6 observers
Detrano et al., 1988 (5, 6)	Review	Exercise thallium scintigraphy	56 primary studies
Detrano et al., 1989 (7)	Review	Exercise electrocardiography	60 primary studies
Doubilet et al., 1981 (52)	Experimental	Radiography	8 test radiographs; 4 with suggestive and 4 nonsuggestive history
Eldevik et al., 1982 (53)	Experimental	Myelography and computed tomography	107 patients assessed with and without clinical history
Elmore et al., 1994 (54)	Experimental	Mammography	150 mammograms, 10 radiologists
Elmore et al., 1997 (55)	Experimental	Mammography	100 radiographs assessed with and without clinical history
Froelicher et al., 1998 (28)	Diagnostic accuracy	Electrocardiography and angiographic calipers	814 consecutive patients with angina
Good et al., 1990 (56)	Experimental	Chest radiography	247 radiographs assessed with and without clinical history
Lijmer et al., 1999 (12)	Review	Various tests	184 primary studies of 218 tests
Potchen et al., 1979 (57)	Experimental	Chest radiography	3 groups of radiologists; different combinations of data
Raab et al., 1995 (58)	Experimental	Bronchial brush specimens	100 bronchial brush specimens examined by different observers
Raab et al., 2000 (59)	Experimental	Bronchial brush specimens	97 specimens, assessed with and without clinical information
Ransohoff et al., 1978 (19)	Review	Carcinoembryonic antigen and nitroblue tetrazolium tests	17 studies of carcinoembryonic antigen and 16 of nitroblue tetrazolium
Ronco et al., 1996 (60)	Experimental	Colpohistologic and cytologic screening	61 samples examined by cytologists and experts
Schreiber, 1963 (61)	Experimental	Chest radiography	100 chest radiographs assessed with and without clinical information

* DSM-III = *Diagnostic and Statistical Manual of the American Psychological Association*, 3rd edition; NA = not applicable; RDC = Research Diagnostic Criteria; ↑ = increased; ↓ = decreased.

tests for acute diseases that may be easily treated (for example, infections) and that may change more rapidly than chronic conditions that do not respond well to treatment and that may remain in the same stage for longer periods.

Reference Standard

The evidence was strong for the effect of biases associated with verification procedure on test performance. All studies that looked at the effects of using an inappropriate reference standard found that test performance was affected; however, the direction of the effect differed among studies. Theoretically, if the reference standard is not 100% accurate, the index test may correctly classify results that have been incorrectly classified by the reference standard. This would be expected to lead to an underestimation of test performance. It is also possible that an imperfect reference standard may classify results of the index test

as being correct when they are actually incorrect. This would be expected to lead to overestimation of test performance. Thus, an inaccurate reference standard could affect test performance in either way.

Many studies looked at the effects of verification bias, especially partial verification bias. Most reported that verification influenced estimates of test performance. In theory, if all of the patients with negative test results are not verified by the reference standard and are subsequently omitted from the 2×2 table, estimates of sensitivity would be inflated because patients with false-negative test results will go undetected. This is supported by the evidence; all studies that observed a significant effect on sensitivity found that sensitivity was increased in the presence of verification bias. However, as with many other biases, the effects on specificity were less clear.

Table 5—Continued

Source of Bias or Variation	Factors Investigated	Effect on Sensitivity	Effect on Specificity	Effect on Overall Accuracy
Inappropriate reference standard	DSM-III instead of RDC as the reference standard	↓	Not reported	NA
Clinical review bias	Availability of clinical information	↑	None	↑
Observer variation	Difference between radiologists and orthopedic surgeons	NA	NA	Associated
Observer variation	Difference between radiologists and orthopedic surgeons	NA	NA	Associated
Observer variation	Inter- and intraobserver variation	NA	NA	Associated
Observer variation	Effect of training and experience	↑	↑	NA
Observer variation	Inter- and intraobserver variation	NA	NA	None
Observer variation	Interobserver variation	NA	NA	Associated
Review bias	Lack of blinding, that is, presence of review bias	↑	Not reported	↑
Review bias	Lack of blinding, that is, presence of review bias	NA	NA	None
Clinical review bias	Suggestive clinical history	↑	↓	NA
Clinical review bias	Availability of clinical information	↑	↓	NA
Observer variation	Interobserver variation	NA	NA	Associated
Clinical review bias	Availability of clinical information	NA	NA	↑
Clinical review bias	Availability of clinical information	↑	Not reported	NA
Clinical review bias	Availability of clinical information	NA	NA	None
Review bias	Lack of blinding, that is, presence of review bias	NA	NA	↑
Clinical review bias	Availability of clinical information	↑	Not reported	NA
Observer variation	Interobserver variation	NA	NA	Associated
Clinical review bias	Availability of clinical information	NA	NA	↑
Review bias	Lack of blinding, that is, presence of review bias	↑	↑	NA
Observer variation	Effect of training and experience (being an "expert")	↑	Not reported	NA
Clinical review bias	Availability of clinical information	↑	None	NA

Interpretation

Reading processes that involve interpretation of results affect estimates of test performance. Both diagnostic and test review biases were found to increase sensitivity; however, no effect on specificity was noted. An effect on sensitivity would be expected because knowledge of the index test result when interpreting the reference standard (or vice versa) probably increases the agreement between tests. This in turn leads to a greater number of true-positives and true-negative results and would be expected to increase estimates of both sensitivity and specificity. It is unclear why studies did not find significant effects on specificity. Perhaps the effects on specificity are smaller and any effect may therefore not reach statistical significance.

The availability of clinical information to the person interpreting the results of the index test was found to increase sensitivity. Although the evidence for an effect on

specificity was minimal, specificity decreased in 2 studies. The provision of clinical information probably has different effects depending on the test being evaluated. Whether clinical information should be available in a particular diagnostic study should be carefully considered in each case. It seems that the best approach to interpreting the results of a diagnostic accuracy study would be to determine whether the clinical information available to those interpreting the results of the index test is the same as the clinical information that would be available when the test is interpreted in practice.

All studies that looked at the effects of observer variation found significant differences among observers in their estimates of test performance. Therefore, the effects of observer variation will inevitably be greater for tests that involve a strong degree of subjective interpretation compared with a fully automated test.

Table 6. Analysis*

Study Details	Design	Index Test	Study Sample	Source of Bias or Variation	Factors Investigated	Effect on Sensitivity	Effect on Specificity	Effect on Overall Accuracy
Detrano et al., 1989 (7)	Review	Exercise electrocardiography	60 primary studies	Handling of indeterminate results	Treatment of equivocal or nondiagnostic tests	NA	NA	None
Philbrick et al., 1982 (18)	Diagnostic accuracy	Graded exercise test	208 consecutive patients	Handling of indeterminate results	Exclusion of unsatisfactory exercise test results	NA	NA	Unclear

* NA = not applicable.

Analysis

Very few studies investigated the effects of biases associated with analysis on test performance. The effect of the exclusion of indeterminate results and the nonarbitrary choice of threshold value remains unclear from the evidence reviewed.

Limitations

The main limitation of our review is the difficulty in identifying articles that examined specific features of the design and conduct of diagnostic studies. Indexing on MEDLINE and other electronic databases focuses on diseases, therapies, and test technologies and not on elements of design. There is no specific way of indexing studies that relate to the diagnostic accuracy of a test (1). In addition, many different names have been used to label the same phenomenon in studies of diagnostic accuracy tests. To try to overcome these difficulties, very broad searches were performed. However, we may have still missed several relevant papers. The information provided in our paper should provide useful examples but may not be comprehensive.

Ideally, we would have liked to provide a quantitative synthesis to assess the magnitude of each of the biases and sources of variation as well as their direction. However, because the studies included were very heterogeneous, a quantitative synthesis was not possible. The studies also measured the effect of the biases and sources of variation in different ways. In particular, diagnostic accuracy and experimental studies looked at the effect of biases and sources of variation within studies, whereas reviews looked at reasons for differences in estimates among studies. It is also likely that different biases and sources of variation will be important in different topic areas. For example, observer variation is likely to be a problem only for studies that involve some degree of subjective interpretation. Also, observer variation is likely to have a greater effect with more subjective interpretations.

Another problem is that sources of bias and variation may act differently depending on the study. For example, for partial verification bias, the effects may differ when the reference standard is not used in selected groups. The group that does not receive verification may, for example, be a random sample of patients, a selected subgroup of patients with negative test results, or all patients with pos-

itive test results. All of these situations are called partial verification, but the effects of each situation probably differ. Within a single study, there is only one true effect of a feature, but this true effect may differ depending on the study. Chance and the effect of other factors may obscure the true effect. These factors combine to create difficulty in determining the overall effect of a source of bias or variation.

We included studies that provided both real-life examples of the effects of different biases and sources of variation as well as studies that used modeling to investigate the effects of different biases or sources of variation. When the results of the modeling studies are interpreted, it is important to consider that these studies can provide an indication only of the theoretical effect of a source of bias or variation. The results from these studies need to be supported by additional empirical evidence from real-life examples before more firm conclusions can be drawn (12).

CONCLUSIONS

This paper provides information on the available evidence for the effects of each source of bias and variation in diagnostic accuracy studies. The sources of bias and variation for which there is the most evidence are demographic features, disease prevalence or severity, partial verification bias, clinical review bias, and observer or instrument variation. Some evidence was also available for the effects of distorted selection of participants, absent or inappropriate reference standard, differential verification bias, and review bias. The potential effects of these biases and sources of variation should be considered when interpreting or designing diagnostic accuracy studies. Additional research should be done to investigate potential sources of bias and variation.

From the University of York, York, United Kingdom, and the University of Amsterdam, Amsterdam, the Netherlands.

Disclaimer: The views expressed in this paper are those of the authors and not necessarily those of the Standing Group, the Commissioning Group, or the Department of Health.

Acknowledgments: The authors thank Kath Wright (Centre for Reviews and Dissemination) for carrying out literature searches. They also

thank the advisory panel to the review for their help during various stages, including commenting on the protocol and draft report.

Grant Support: Commissioned and funded by the National Health Service R&D Health Technology Assessment Programme (project number 98/27/99).

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Penny Whiting, MSc, Centre for Reviews and Dissemination, University of York, York YO10 5DD, United Kingdom; e-mail, pfw2@york.ac.uk.

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

References

- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem*. 2003;49:7-18. [PMID: 12507954]
- Feinstein AR. *Clinical Epidemiology: The Architecture of Clinical Research*. Philadelphia: WB Saunders Co; 1985.
- Whiting PJ, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. The development and validation of methods for assessing the quality and reporting of diagnostic studies. *Health Technol Assess* [In press].
- Curtin F, Morabia A, Pichard C, Slosman DO. Body mass index compared to dual-energy x-ray absorptiometry: evidence for a spectrum bias. *J Clin Epidemiol*. 1997;50:837-43. [PMID: 9253396]
- Detrano R, Janosi A, Lyons KP, Marcondes G, Abbassi N, Froelicher VF. Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. *Am J Med*. 1988;84:699-710. [PMID: 3041808]
- Detrano R, Lyons KP, Marcondes G, Abbassi N, Froelicher VF, Janosi A. Methodologic problems in exercise testing research. Are we solving them? *Arch Intern Med*. 1988;148:1289-95. [PMID: 3288157]
- Detrano R, Gianrossi R, Mulvihill D, Lehmann K, Dubach P, Colombo A, et al. Exercise-induced ST segment depression in the diagnosis of multivessel coronary disease: a meta analysis. *J Am Coll Cardiol*. 1989;14:1501-8. [PMID: 2809010]
- Eggin TK, Feinstein AR. Context bias. A problem in diagnostic radiology. *JAMA*. 1996;276:1752-5. [PMID: 8940325]
- Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. *Am J Med*. 1984;77:64-71. [PMID: 6741986]
- Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med*. 1992;117:135-40. [PMID: 1605428]
- Levy D, Labib SB, Anderson KM, Christiansen JC, Kannel WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. *Circulation*. 1990;81:815-20. [PMID: 2137733]
- Lijmer JG, Mol BW, Heisterkamp S, Bossel 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]
- Melbye H, Straume B. The spectrum of patients strongly influences the usefulness of diagnostic tests for pneumonia. *Scand J Prim Health Care*. 1993; 11:241-6. [PMID: 8146507]
- Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology*. 1997;8:12-7. [PMID: 9116087]
- Morise AP, Diamond GA. Does sex discrimination explain the differences in test accuracy among men and women referred for exercise electrocardiography? *Circulation*. 1994;90(Pt 2):1-273.
- Morise AP, Diamond GA. Comparison of the sensitivity and specificity of exercise electrocardiography in biased and unbiased populations of men and women. *Am Heart J*. 1995;130:741-7. [PMID: 7572581]
- O'Connor PW, Tansay CM, Detsky AS, Mushlin AI, Kucharczyk W. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. *Neurology*. 1996;47: 140-4. [PMID: 8710067]
- Philbrick JT, Horwitz RI, Feinstein AR, Langou RA, Chandler JP. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. *JAMA*. 1982;248:2467-70. [PMID: 7131702]
- Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926-30. [PMID: 692598]
- Roger VL, Pellikka PA, Bell MR, Chow CW, Bailey KR, Seward JB. Sex and test verification bias. Impact on the diagnostic value of exercise echocardiography. *Circulation*. 1997;95:405-10. [PMID: 9008457]
- Rozanski A, Diamond GA, Berman D, Forrester JS, Morris D, Swan HJ. The declining specificity of exercise radionuclide ventriculography. *N Engl J Med*. 1983;309:518-22. [PMID: 6877322]
- Santana-Boado C, Candell-Riera J, Castell-Conesa J, Aguada-Bruix S, Garcia-Burillo A, Canela T, et al. Diagnostic accuracy of technetium-99m-MIBI myocardial SPECT in women and men. *J Nucl Med*. 1998;39:751-5. [PMID: 9591568]
- Stein PD, Gottschalk A, Henry JW, Shivkumar K. Stratification of patients according to prior cardiopulmonary disease and probability assessment based on the number of mismatched segmental equivalent perfusion defects. Approaches to strengthen the diagnostic value of ventilation/perfusion lung scans in acute pulmonary embolism. *Chest*. 1993;104:1461-7. [PMID: 8222807]
- Steinbauer JR, Cantor SB, Holzer CE 3rd, Volk RJ. Ethnic and sex bias in primary care screening tests for alcohol use disorders. *Ann Intern Med*. 1998;129: 353-62. [PMID: 9735062]
- Taube A, Tholander B. Over- and underestimation of the sensitivity of a diagnostic malignancy test due to various selections of the study population. *Acta Oncol*. 1990;29:971-6. [PMID: 2278729]
- van der Schouw YT, Van Dijk R, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. *J Clin Epidemiol*. 1995;48:417-22. [PMID: 7897462]
- van Rijkom HM, Verdonschot EH. Factors involved in validity measurements of diagnostic tests for approximal caries—a meta-analysis. *Caries Res*. 1995;29:364-70. [PMID: 8521438]
- Froelicher VF, Lehmann KG, Thomas R, Goldman S, Morrison D, Edson R, et al. The electrocardiographic exercise test in a population with reduced workup bias: diagnostic performance, computerized interpretation, and multivariable prediction. Veterans Affairs Cooperative Study in Health Services #016 (QUEXTA) Study Group. Quantitative Exercise Testing and Angiography. *Ann Intern Med*. 1998;128:965-74. [PMID: 9625682]
- Arana GW, Zarzar MN, Baker E. The effect of diagnostic methodology on the sensitivity of the TRH stimulation test for depression: a literature review. *Biol Psychiatry*. 1990;28:733-7. [PMID: 2122917]
- Bowler JV, Munoz DG, Merskey H, Hachinski V. Fallacies in the pathological confirmation of the diagnosis of Alzheimer's disease. *J Neurol Neurosurg Psychiatry*. 1998;64:18-24. [PMID: 9436722]
- Boyko EJ, Alderman BW, Baron AE. Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. *J Gen Intern Med*. 1988;3:476-81. [PMID: 3049969]
- Cecil MP, Kosinski AS, Jones MT, Taylor A, Alazraki NP, Pettigrew RI, et al. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. *J Clin Epidemiol*. 1996;49:735-42. [PMID: 8691222]
- Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39:207-15. [PMID: 6871349]
- De Neef P. Evaluating rapid tests for streptococcal pharyngitis: the apparent accuracy of a diagnostic test when there are errors in the standard of comparison. *Med Decis Making*. 1987;7:92-6. [PMID: 3553828]
- Diamond GA. Affirmative actions: can the discriminant accuracy of a test be determined in the face of selection bias? *Med Decis Making*. 1991;11:48-56. [PMID: 2034075]
- Diamond GA. Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem. *Med Decis Making*. 1992;12:22-31. [PMID: 1160877]

1538629]

37. Lijmer JG, Hunink MG, van den Dungen JJ, Loonstra J, Smit AJ. ROC analysis of noninvasive tests for peripheral arterial disease. *Ultrasound Med Biol*. 1996;22:391-8. [PMID: 8795165]
38. Miller TD, Hodge DO, Christian TF, Milavetz JJ, Bailey KR, Gibbons RJ. The impact of adjusting for post-test referral bias on apparent sensitivity and specificity of SPECT myocardial perfusion imaging in men and women [Abstract]. *J Am Coll Cardiol*. 1998;31(2 Suppl A):167A.
39. Mol BW, Lijmer JG, van der Meulen J, Pajkt E, Bilardo CM, Bossuyt PM. Effect of study design on the association between nuchal translucency measurement and Down syndrome. *Obstet Gynecol*. 1999;94:864-9. [PMID: 10546775]
40. Panzer RJ, Suchman AL, Griner PF. Workup bias in prediction research. *Med Decis Making*. 1987;7:115-9. [PMID: 3574021]
41. Phelps CE, Hutson A. Estimating diagnostic test accuracy using a "fuzzy gold standard." *Med Decis Making*. 1995;15:44-57. [PMID: 7898298]
42. Ransohoff DF, Muir WA. Diagnostic workup bias in the evaluation of a test. Serum ferritin and hereditary hemochromatosis. *Med Decis Making*. 1982;2:139-45. [PMID: 7167042]
43. Thibodeau L. Evaluating diagnostic tests. *Biometrics*. 1981:801-4.
44. Zhou XH. Effect of verification bias on positive and negative predictive values. *Stat Med*. 1994;13:1737-45. [PMID: 7997707]
45. Detrano R, Gianrossi R, Froelicher V. The diagnostic accuracy of the exercise electrocardiogram: a meta-analysis of 22 years of research. *Prog Cardiovasc Dis*. 1989;32:173-206. [PMID: 2530605]
46. Berbaum KS, el-Khoury GY, Franken EA Jr, Kathol M, Montgomery WJ, Hesson W. Impact of clinical history on fracture detection with radiography. *Radiology*. 1988;168:507-11. [PMID: 3393672]
47. Berbaum KS, Franken EA Jr, el-Khoury GY. Impact of clinical history on radiographic detection of fractures: a comparison of radiologists and orthopedists. *AJR Am J Roentgenol*. 1989;153:1221-4. [PMID: 2816635]
48. Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intra-observer variability of mammogram interpretation: a field study. *Eur J Cancer*. 1992;28A:1054-8. [PMID: 1627374]
49. Cohen MB, Rodgers RP, Hales MS, Gonzales JM, Ljung BM, Beckstead JH, et al. Influence of training and experience in fine-needle aspiration biopsy of breast. Receiver operating characteristics curve analysis. *Arch Pathol Lab Med*. 1987;111:518-20. [PMID: 3579506]
50. Corley DE, Kirtland SH, Winterbauer RH, Hammar SP, Dail DH, Bauermeister DE, et al. Reproducibility of the histologic diagnosis of pneumonia among a panel of four pathologists: analysis of a gold standard. *Chest*. 1997;112:458-65. [PMID: 9266884]
51. Cuaron A, Acero AP, Cardenas M, Huerta D, Rodriguez A, de Garay R. Interobserver variability in the interpretation of myocardial images with Tc-99m-labeled diphosphonate and pyrophosphate. *J Nucl Med*. 1980;21:1-9. [PMID: 6243350]
52. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. *AJR Am J Roentgenol*. 1981;137:1055-8. [PMID: 6975000]
53. Eldevik OP, Dugstad G, Orrison WW, Haughton VM. The effect of clinical bias on the interpretation of myelography and spinal computed tomography. *Radiology*. 1982;145:85-9. [PMID: 7122902]
54. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med*. 1994;331:1493-9. [PMID: 7969300]
55. Elmore JG, Wells CK, Howard DH, Feinstein AR. The impact of clinical history on mammographic interpretations. *JAMA*. 1997;277:49-52. [PMID: 8980210]
56. Good BC, Cooperstein LA, DeMarino GB, Miketic LM, Gennari RC, Rockette HE, et al. Does knowledge of the clinical history affect the accuracy of chest radiograph interpretation? *AJR Am J Roentgenol*. 1990;154:709-12. [PMID: 2107662]
57. Potchen E, Gard J, Lazar P, Lahaie P, Andary M. The effect of clinical history data on chest film interpretation: direction or distraction [Abstract]. *Invest Radiol*. 1979;14:404.
58. Raab SS, Thomas PA, Lenel JC, Bottles K, Fitzsimmons KM, Zaleski MS, et al. Pathology and probability. Likelihood ratios and receiver operating characteristic curves in the interpretation of bronchial brush specimens. *Am J Clin Pathol*. 1995;103:588-93. [PMID: 7741104]
59. Raab SS, Oweity T, Hughes JH, Salomao DR, Kelley CM, Flynn CM, et al. Effect of clinical history on diagnostic accuracy in the cytologic interpretation of bronchial brush specimens. *Am J Clin Pathol*. 2000;114:78-83. [PMID: 10884802]
60. Ronco G, Montanari G, Aimone V, Parisio F, Segnan N, Valle A, et al. Estimating the sensitivity of cervical cytology: errors of interpretation and test limitations. *Cytopathology*. 1996;7:151-8. [PMID: 8782987]
61. Schreiber M. The clinical history as a factor in roentgenogram interpretation. *JAMA*. 1963;185:137-9.

Current Author Addresses: Ms. Whiting and Dr. Kleijnen: Centre for Reviews and Dissemination, University of York, York YO10 5DD, United Kingdom.

Ms. Rutjes and Drs. Reitsma, Glas, and Bossuyt: Department of Clinical Epidemiology & Biostatistics, Academic Medical Center, University of Amsterdam, P.O. Box 22700, 1100 DE Amsterdam, the Netherlands.