American Journal of Epidemiology Advance Access published March 15, 2012

American Journal of Epidemiology

DOI: 10.1093/aje/kwr383

Practice of Epidemiology

Adjusting for Partial Verification or Workup Bias in Meta-Analyses of Diagnostic Accuracy Studies

Joris A. H. de Groot*, Nandini Dendukuri, Kristel J. M. Janssen, Johannes B. Reitsma, James Brophy, Lawrence Joseph, Patrick M. M. Bossuyt, and Karel G. M. Moons

* Correspondence to Dr. Joris A. H. de Groot, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508GA Utrecht, the Netherlands (e-mail: j.degroot-17@umcutrecht.nl).

Initially submitted October 22, 2010; accepted for publication September 23, 2011.

A key requirement in the design of diagnostic accuracy studies is that all study participants receive both the test under evaluation and the reference standard test. For a variety of practical and ethical reasons, sometimes only a proportion of patients receive the reference standard, which can bias the accuracy estimates. Numerous methods have been described for correcting this partial verification bias or workup bias in individual studies. In this article, the authors describe a Bayesian method for obtaining adjusted results from a diagnostic meta-analysis when partial verification or workup bias is present in a subset of the primary studies. The method corrects for verification bias without having to exclude primary studies with verification bias, thus preserving the main advantages of a meta-analysis: increased precision and better generalizability. The results of this method are compared with the existing methods for dealing with verification bias in diagnostic meta-analyses. For illustration, the authors use empirical data from a systematic review of studies of the accuracy of the immunohistochemistry test for diagnosis of human epidermal growth factor receptor 2 status in breast cancer patients.

diagnosis; meta-analysis; verification bias; workup bias

Abbreviations: CrI, credible interval; FISH, fluorescence in situ hybridization; HER-2, human epidermal growth factor receptor 2; IHC, immunohistochemistry.

An increasing number of systematic reviews of diagnostic accuracy studies are being published. These studies aim to provide more precise and more generalizable accuracy estimates and to examine variability in accuracy across clinical subgroups in a more meaningful way than can be done in separate, small studies (1). Systematic reviews of test accuracy studies have benefited from methodological advances and guidelines for the design and interpretation of primary diagnostic studies (2–6). Methods for meta-analysis now include bivariate or hierarchical models, which jointly summarize sensitivity and specificity while accounting for their mutual relation within and across primary studies. Unlike most meta-analyses of therapeutic trials, which are usually based on randomized trial data, diagnostic meta-analyses often involve primary studies based on routinely collected data. Therefore, primary studies in a diagnostic meta-analysis may be more susceptible to a number of well-documented sources of bias, such as selection or misclassification bias (7, 8). So far, few authors have tried to correct for biases within a diagnostic meta-analysis, although some have attempted to correct for bias from an imperfect reference standard using a latent class model (9–11).

One of the most problematic biases in primary diagnostic accuracy studies is perhaps the so-called selection, workup, or verification bias (8, 12–14). A classical scenario in which this bias arises is a 2-stage design, where all subjects undergo the test under evaluation or the index test at stage I but only a sample of subjects are selected at stage II to undergo verification of disease presence by the reference standard. When selection of subjects for the reference standard is not completely random, verification bias will occur. This could happen when, for example, a stratified random sample is drawn in stage II with the strata being defined by the results of the index test in stage I. Such a nonrandom referral pattern may

arise because of ethical or economic considerations—for example, in cases of a low disease probability in index-test-negative subjects—or because of the invasiveness or costs associated with the reference standard.

Several methods exist for addressing this particular form of workup or partial verification bias in primary studies (13, 15, 16). So far, these solutions have not been applied to or developed for systematic reviews and meta-analyses, where partial verification is present in one or more of the primary studies. Some review authors have acknowledged this bias in their discussions but have not quantified or corrected for it in the analyses (17–19). In a recent article, Chu et al. (11) described a meta-analytical method for the case in which primary diagnostic studies have missing data on the reference standard, but the authors did not explicitly address the problem as partial verification, workup, or selection bias.

Excluding all primary studies with workup bias is one simple and frequently applied solution (2). This method avoids the partial verification bias, but at the expense of reduced precision and lower generalizability. Alternatively, sensitivity analyses that exclude the questionable studies may be performed, to assess the robustness of the conclusions (2, 20). However, both methods can leave the researcher with questionable results, since omitting studies results in possible publication bias, and including them may result in verification bias. It would be preferable to include all studies and adjust for the verification bias that may be present.

Here we extend existing methods of correcting for partial verification bias in single diagnostic studies (13) to the meta-analytic setting. In particular, we propose a 2-stage Bayesian approach to correct for verification bias in primary diagnostic accuracy studies, when conducting a meta-analysis of test accuracy studies. In stage I of the analysis, this approach uses only the unbiased primary studies to estimate the distribution of the index test results in a representative sample of the population. In stage II, all available studies are used to estimate positive predictive values. The results from the 2 stages can then be combined to obtain unbiased summary estimates of the sensitivity and specificity of the index test.

EXAMPLE STUDY: HER-2-POSITIVE BREAST CANCER

We will illustrate our method using data from a recently published systematic review on testing for human epidermal growth factor receptor 2 (HER-2)-positive breast cancer, an aggressive form of breast cancer associated with a high mortality rate (21). The availability of Herceptin (Genentech, Inc., San Francisco, California), an effective but expensive treatment for HER-2-positive breast cancer, has increased awareness about the need to accurately identify women who have HER-2 receptors and thus are most likely to respond to this therapy. Two tests are commonly used to determine HER-2 status: immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH). FISH is believed to be the gold standard test for determining HER-2 status. The technique is carried out only at specialized laboratories (21). IHC, on the other hand, can be performed in most surgical pathology laboratories and is substantially less expensive than FISH (21). The goal of the systematic review was to obtain summary estimates of the sensitivity and specificity of IHC, assuming FISH to be a perfect reference standard, and to subsequently compare the cost-effectiveness of different strategies for establishing HER-2 status.

The IHC test is scored on a 4-point scale and takes the values 0, 1+, 2+, or 3+. Patients who receive scores of 0 or 1+ are considered HER-2-negative, while those with scores of 3+ are considered HER-2-positive. Patients with a score of 2+ are considered to have an ambiguous test result. Various authors have recommended that the results of patients with IHC scores of 2+ be verified with a FISH test (22). In more recent studies (21, 23), investigators have recommended that patients with IHC scores of 3+ also have their HER-2 status verified by the FISH test. This implies that patients who receive IHC scores of either 2+ or 3+ are more likely to have their HER-2 status verified by FISH in routine clinical practice than are those who receive scores of 0 or 1+. In our analyses, we treated the IHC as having a 3-point scale: 0 or 1+, 2+, and 3+. The FISH test gives a dichotomous test result of positive or negative. The percentage of HER-2-positive cases in a representative sample of women diagnosed with breast cancer is believed to be approximately 30% (21). Thus, we would expect that studies with a workup bias have a proportion of IHC 2+ and 3+ scores greater than 30%.

Characteristics of participants in the 17 studies included in this meta-analysis (24–40) are summarized in Table 1. Eight of the studies were considered to have partial verification bias resulting in an overrepresentation of cases with 2+ or 3+ IHC scores (33–40). In 4 of these studies, it was evident from the published methods that the study sample was selected at a center where patients had been selectively referred for FISH (36, 38–40). In one study, the study design involved oversampling of patients with IHC results of 2+ (37). We treated an additional 3 studies as having verification bias, even though this was not clear from the articles themselves, because the percentage of 2+ or 3+ cases was higher than 40% (33–35).

In studies without verification bias, the percentage of patients in the IHC 0 or 1+ category ranged from 63% to 85%, as compared with 2.8%–57% in the studies that were considered to have verification bias.

MATERIALS AND METHODS

We first present the general concept behind our method, followed by more mathematical details.

General approach

We assumed that all primary studies had a 2-stage data collection process. At stage I, a random sample of subjects was selected to undergo the index test. At stage II, a certain percentage of these subjects had their HER-2 status verified by means of the reference standard. In primary studies without verification bias, 100% of patients who received the index test at stage I went on to receive the reference standard test at stage II. However, in those studies where there was verification bias, an unknown proportion of patients received the reference standard at stage II. In studies with verification bias, researchers would typically report only results based on the subset of patients evaluated in stage II.

Table 1. Studies Included in a Meta-Analysis of the Sensitivity and Specificity of Immunohistochemistry With Respect to Fluorescence In Situ Hybridization for Diagnosis of HER-2 Breast Cancer

First Author, Year (Reference No.)	No. of Patients	IHC Score, % of Patients			% of Patients With Positive FISH Result in Each IHC Score Category		
		0 or 1+	2+	3+	0 or 1+	2+	3+
Hoang, 2000 (24)	100	74.0	2.0	24.0	0.0	0.0	70.8
Kakar, 2000 (25)	112	70.5	15.2	14.3	1.3	35.2	87.5
Bartlett, 2001 (26)	210	85.2	10.0	4.8	6.7	90.5	90.0
Tsuda, 2001 (27)	101	76.3	5.9	17.8	2.6	0.0	83.3
Press, 2002 (28)	117	74.4	11.1	14.5	14.9	100.0	100.0
Dowsett, 2003 (29)	426	63.4	12.7	23.9	0.7	48.1	94.1
Ogura, 2003 (30)	110	71.9	9.1	18.2	3.7	10.0	100.0
Lal, 2004 (31)	2,279	76.0	13.7	10.3	1.9	26.5	89.7
Lottner, 2005 (32)	215	78.1	11.6	10.2	2.4	72.0	100.0
Lebeau, 2001 (34)[a]	78	56.4	20.5	23.1	0.0	25.0	100.0
McCormick, 2002 (35)[a]	198	56.6	22.7	20.7	6.3	42.3	100.0
Roche, 2002 (36)[a]	119	16.0	10.1	73.9	0.0	0.0	89.8
Mrozkowiak, 2004 (37)[a]	360	2.8	87.5	9.7	0.0	20.3	91.4
Yaziji, 2004 (38)[a]	2,913	49.0	39.5	11.5	2.8	17.0	91.6
Dolan, 2005 (39)[a]	129	17.9	72.1	10.1	0.0	7.5	38.4
Press, 2005 (40)[a]	842	54.3	14.7	31.0	4.2	16.9	78.2
Loring, 2005 (33)[a]	110	56.4	15.5	28.2	0.0	0.0	87.1

Abbreviations: FISH, fluorescence in situ hybridization; HER-2, human epidermal growth factor receptor 2; IHC, immunohistochemistry.
[a] Study with verification bias.

The meta-analysis was also carried out in 2 stages to reflect the 2-stage data collection process. In the first stage of the meta-analysis, we estimated the probability distribution (i.e., the prevalence of each value) of the index test using the primary studies without verification bias. In the second stage of the meta-analysis, we estimated the positive predictive values of the index test across all primary studies, irrespective of whether they had verification bias or not. Following Begg and Greenes (13), we assumed that the estimate of the positive predictive value, $P(\text{Reference} + |\text{Index})$, in each study remained unbiased even in the presence of verification bias. We used a Bayesian approach to estimate the parameters in each stage of the meta-analysis. A WinBUGS program for implementing the model is given in the Web Appendix (http://aje.oxfordjournals.org/). Finally, the pooled sensitivity and specificity of the index test were obtained as functions of the parameters estimated in the 2 stages of the meta-analysis.

Stage I distribution of (index) test results

We assume that the index test results are expressed on an ordinal scale, while the reference standard test results are dichotomous. Let $(t_{1j}, t_{2j}, \ldots, t_{Ij})$ denote the number of subjects in the jth study with results $1, \ldots, I$, respectively, on the index test T. We assume that the vector of index test results follows a multinomial distribution with probability vector $(p_{1j}, \ldots, p_{Ij})$ and sample size $n_j = t_{1j} + t_{2j} + \ldots + t_{Ij}$. Following the approach commonly used to model a receiver operating characteristic curve, we assume that each multinomial probability can be expressed as a difference between 2 cumulative probabilities, $p_{ij} = q_{ij} - q_{i-1j}$ (41). Each q_{ij} can be expressed as a probit (cumulative normal probability) function of a continuous variable a_{ij}, that is, $q_{ij} = \Phi(a_{ij})$. This transformation makes it easier to define a hierarchical prior distribution for the multinomial probabilities. The a_{ij} are assumed to be a random sample from a truncated normal distribution $N(A_i, \sigma_i)$, $a_{i\text{-}1j} \leq a_{ij} \leq a_{i+1j}$, where A_i denotes the pooled mean value of the a_{ij}'s across studies and σ_i is the between-study standard deviation. The truncated distribution helps preserve the ordering among the a_{ij}'s. For each study, q_{0j} is assumed to be 0 and q_{Ij} is assumed to be 1. The lower limit of truncation for a_{1j} is $-\infty$, and the upper limit of truncation for a_{Ij} is ∞. We used objective $N(\text{mean} = 0, \text{standard deviation} = 10)$ and uniform(0, 100) prior distributions for each of A_i and σ_i, respectively.

Stage II distribution of reference standard results

We assume that in the jth study, we observe the variables r_{ij}, $i = 1, \ldots, I$ denoting the number of subjects with a positive result on the reference standard given the result $T = i$ on the index test. We assume that each r_{ij} follows a binomial distribution with probability s_{ij} and sample size t_{ij}. The probabilities s_{ij} are expressed as a probit function of a continuous variable b_{ij}, that is, $s_{ij} = \Phi(b_{ij})$. The b_{ij}'s are assumed to follow a normal distribution $N(B_i, \tau_i)$, where B_i is the pooled mean of the b_{ij}'s across all studies and τ_i is the between-study standard deviation. Once again, objective prior distributions

Table 2. Overall Results of Meta-Analysis for the Percentage of Patients in Each Immunohistochemistry Category (Stage I) and the Percentage of Patients With a Positive Fluorescence In Situ Hybridization Test Result in Each Immunohistochemistry Category (Stage II)

	IHC Score					
	0 or 1+		2+		3+	
	Median	95% CrI	Median	95% CrI	Median	95% CrI
IHC score, probability (stage I)	0.77	0.73, 0.80	0.11	0.05, 0.18	0.13	0.09, 0.18
Probability of a positive FISH result in each IHC score category (stage II)	0.03	0.02, 0.04	0.27	0.11, 0.48	0.91	0.85, 0.95

Abbreviations: CrI, credible interval; FISH, fluorescence in situ hybridization; IHC, immunohistochemistry.

are normal with mean = 0 and standard deviation = 10 and uniform(0, 100) for each of B_i and τ_i, respectively.

Obtaining a sample from the posterior distribution

Neither in stage I nor in stage II can the posterior distribution be expressed in a simple analytical form. Using a WinBUGS program, we obtained a sample from the posterior distribution of each parameter of interest via Markov chain Monte Carlo methods. For each model described in this paper, 5 Markov chain Monte Carlo runs were carried out with different starting values. Convergence of the model was determined using the Gelman-Rubin statistic provided by WinBUGS (42). Once model convergence was ascertained, we drew a sample of 500,000 iterations after dropping the first 10,000 burn-in iterations. This sample was used to obtain summary statistics (e.g., median value and 2.5% and 97.5% quantiles).

Estimating the pooled sensitivity and specificity of the index test

Let $P_1 = \Phi(A_1), P_2 = \Phi(A_2) - \Phi(A_1), \ldots, P_I = 1 - \Phi(\mathrm{A}_{I-1})$ denote the pooled estimates across all studies of the prevalence of each value of the index test. Similarly, let $S_1 = \Phi(B_1)$, $S_2 = \Phi(B_2), \ldots, S_I = \Phi(B_I)$ denote the pooled estimates of the probability of a positive result on the reference standard for a given result of the index test.

The sensitivity of the index test at the cutoff of $T = i$ can be defined as

$$\frac{\sum_{k=i}^{I} S_k P_k}{\sum_{k=1}^{I} S_k P_k}.$$

Similarly, the specificity at the cutoff of $T = i$ can be defined as

$$\frac{\sum_{k=1}^{i-1} (1 - S_k) P_k}{\sum_{k=1}^{I} (1 - S_k) P_k}.$$

Sensitivity analyses

We carried out a sensitivity analysis by considering a lower cutoff for defining verification bias $P(\mathrm{IHC} = 2+ \text{ or } 3+) > 30\%$. This would imply that the study by Dowsett et al. (29) would also be considered to have verification bias and would be included only in stage II of the meta-analysis.

Selecting the prior distribution for the parameters modeling between-study heterogeneity in a hierarchical model can be potentially problematic. We follow the approach described by Gilks et al. (43) to assess the sensitivity of our inferences to commonly used low information prior distributions. In addition to the uniform(0, 100) prior over the standard deviation, we fit the model with a gamma(0.001, 0.001) prior over the between-study precision and a uniform(0, 100) prior over the between-study variance.

Comparison with other methods of adjusting for verification bias in a meta-analysis

We compared the results of the 2-stage model described above with the results obtained when 1) verification bias was ignored and 2) studies with verification bias were excluded in total from the analysis.

RESULTS

Distribution of index test and reference standard results

The first row of Table 2 shows the posterior estimates (median values and 95% credible intervals) derived from stage I of the model, for the probability of each value of the IHC test. The second row of Table 2 shows the posterior estimates (median values and 95% credible intervals) derived from stage II of the model, for the probability of a positive FISH test result in each IHC score category.

The wide credible intervals for both the overall IHC scores and the positive FISH results (especially the 2+ and 3+ categories) indicated that there was a substantial amount of between-study variability.

Pooled sensitivities and specificities per correction method

Table 3 lists the pooled estimates of the sensitivity (top) and specificity (bottom) of the IHC test obtained using each of the different methods.

Estimates obtained from the adjusted model and the model that relied only on studies without verification bias were similar, though the precision was worse in the latter case, as expected.

Table 3. Sensitivities and Specificities of the Possible Immunohistochemistry Scores Using the Authors' Bayesian Method and 2 Other Methods: 1) Ignoring Verification Bias in the Analysis and 2) Excluding Studies With Verification Bias in the Analysis

IHC Score Cutoff	Verification Bias Corrected Using the Bayesian Method		Ignoring Verification Bias		Excluding Studies With Verification Bias	
	Median	95% CrI	Median	95% CrI	Median	95% CrI
Sensitivity of IHC at different cutoffs						
≥0	1		1		1	
≥2	0.88	0.82, 0.93	0.94	0.89, 0.97	0.89	0.79, 0.95
≥3	0.72	0.54, 0.86	0.66	0.42, 0.85	0.64	0.43, 0.85
>3	0		0		0	
Specificity of IHC at different cutoffs						
≥0	0		0		0	
≥2	0.89	0.83, 0.95	0.72	0.50, 0.89	0.91	0.85, 0.97
≥3	0.98	0.97, 0.99	0.98	0.96, 0.99	0.99	0.97, 0.99
>3	1		1		1	

Abbreviations: CrI, credible interval; IHC, immunohistochemistry.

At the cutoff of 2+, we found that the sensitivity obtained with the method that ignored verification bias altogether was higher, while specificity was considerably lower, in comparison with the results from the adjusted model. At the cutoff of 3+, the pattern was reversed, with the sensitivity being lower than in the adjusted model. This was probably because the IHC 2+ subgroup was more likely to be oversampled at stage II than the IHC 3+ subgroup.

These results are similar to what has been found when ignoring verification bias in primary diagnostic studies (8, 12–14). That is, when oversampling index-test-positive subjects (IHC ≥2+), the sensitivity is overestimated while the specificity is underestimated, whereas when oversampling index-test-negative patients (IHC ≤2+), the sensitivity is underestimated while the specificity is overestimated.

Sensitivity analysis

Sensitivity analyses, altering the cutoff to classify a study as having verification bias, did not have an important impact on the pooled median values and 95% credible intervals of the sensitivities and specificities (sensitivity at 2+ = 0.88, 95% credible interval (CrI): 0.81, 0.92; sensitivity at 3+ = 0.71, 95% CrI: 0.56, 0.82; specificity at 2+ = 0.90, 95% CrI: 0.85, 0.93; specificity at 3+ = 0.99, 95% CrI: 0.97, 0.99).

Changing the form of the prior distribution for the between-study heterogeneity parameters did not alter the results in Table 3 (results not shown).

DISCUSSION

In this paper, we have presented a method of adjusting for workup bias or partial verification bias in a diagnostic meta-analysis. We compared the results of this method with several alternative approaches, including the naive approach of prevailing methods such as simply ignoring the verification bias in the primary studies.

In our empirical example, it appears that ignoring verification bias results in a bias similar to that observed in individual diagnostic studies with verification bias (8, 12–14). This supports our notion that in a diagnostic meta-analysis context as well, verification bias in primary studies can lead to seriously biased results and should be addressed to make valid inferences about the test under study.

The performance of our method relies on having at least some primary studies without verification bias to be able to correct the primary studies with this bias. In our empirical example, simply omitting studies with verification led to results similar to the corrected values achieved by our Bayesian model. However, leaving studies completely out of the analysis can generally lead to different estimates and will always reduce the overall sample size, leading to lower precision of the parameter estimates associated with the index test. The more primary studies one omits from the analysis (due to verification bias), the more this will affect both bias and precision. Meta-analyses are in principle done to improve precision and generalizability, so leaving studies out of the analysis and therefore completely ignoring valuable information should not be preferred.

An important step before applying the proposed correction method is to identify primary studies with or without partial verification bias. In some studies, the presence of the bias is evident from reading the Methods section in the primary studies. In other situations, however, the bias is not clearly reported, and the presence or absence of verification bias has to be assessed on clinical and methodological grounds. Because this is a more subjective assessment, it may increase the risk of misclassification of studies; therefore, we recommend carrying out a sensitivity analysis similar to that used in our example.

Our model extends the methods of Begg and Greenes (13) that correct for verification bias within a single study to the meta-analysis context. A key assumption in the adjustment is that the predictive values of the index test can be estimated without bias, even in studies with verification bias. The validity of the Begg and Greenes method has been thoroughly studied (13, 44). Therefore, although we illustrate our method using only one example study of partial verification bias in primary diagnostic accuracy studies when conducting a meta-analysis, there is no reason to believe that the properties of the method will not carry over to other settings.

The model proposed here can be extended to incorporate both covariates that influence the distribution of index test results and covariates that influence the predictive values of the index test. As we mentioned in the Introduction, misclassification due to an imperfect reference standard is a well-recognized problem in diagnostic testing studies. As has been described for single studies, we can also extend our model to simultaneously correct for verification bias and bias due to imperfection of the reference standard (45, 46). Should the stage I data be available in some of the studies with verification bias, then we could also add a further step to our model to estimate the probability of verification. This would be particularly important if additional covariates besides the index test results determined the probability of verification and also affected the distribution of the index test and the positive predictive values.

It is well known that verification bias in primary diagnostic accuracy studies, as well as in meta-analysis of such studies, can seriously harm estimates of the diagnostic accuracy of the index test. Our proposed model corrects for this bias without excluding any primary studies with verification bias and thus preserves the main advantages of a meta-analysis: increased precision and better generalizability.

ACKNOWLEDGMENTS

Author affiliations: Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht, the Netherlands (Joris A. H. de Groot, Kristel J. M. Janssen, Karel G. M. Moons); Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, the Netherlands (Johannes B. Reitsma, Patrick M. M. Bossuyt); and Department of Epidemiology and Biostatistics, Faculty of Medicine, McGill University, Montreal, Quebec, Canada (Nandini Dendukuri, James Brophy, Lawrence Joseph).

The authors acknowledge the support of the Netherlands Organization for Scientific Research (projects 9120.8004 and 918.10.615).

Conflict of interest: none declared.

REFERENCES

1. Leeflang MM, Deeks JJ, Gatsonis C, et al. Systematic reviews of diagnostic test accuracy. Cochrane Diagnostic Test Accuracy Working Group. *Ann Intern Med*. 2008; 149(12):889–897.
2. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994;120(8):667–676.
3. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20(19):2865–2884.
4. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*. 2005; 58(10):982–990.
5. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *AJR Am J Roentgenol*. 2006;187(2):271–281.
6. Zhou XH, Brizendine EJ, Pritz MB. Methods for combining rates from several studies. *Stat Med*. 1999;18(5):557–566.
7. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004;140(3):189–202.
8. Rutjes AW, Reitsma JB, Di Nisio M, et al. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006; 174(4):469–476.
9. Sadatsafavi M, Shahidi N, Marra F, et al. A statistical method was used for the meta-analysis of tests for latent TB in the absence of a gold standard, combining random-effect and latent-class methods to estimate test accuracy. *J Clin Epidemiol*. 2010;63(3):257–269.
10. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. *J Clin Epidemiol*. 1999;52(10):943–951.
11. Chu H, Chen S, Louis TA. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. *J Am Stat Assoc.* 2009;104(486):512–523.
12. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299(17):926–930.
13. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39(1):207–215.
14. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282(11):1061–1066.
15. Harel O, Zhou XH. Multiple imputation for correcting verification bias. *Stat Med*. 2006;25(22):3769–3786.
16. de Groot JA, Janssen KJ, Zwinderman AH, et al. Multiple imputation to correct for partial verification bias revisited. *Stat Med*. 2008;27(28):5880–5889.
17. Met R, Bipat S, Legemate DA, et al. Diagnostic performance of computed tomography angiography in peripheral arterial disease: a systematic review and meta-analysis. *JAMA*. 2009; 301(4):415–424.
18. Nayak S, Olkin I, Liu H, et al. Meta-analysis: accuracy of quantitative ultrasound for identifying patients with osteoporosis. *Ann Intern Med*. 2006;144(11):832–841.
19. Mijnhout GS, Hoekstra OS, van Tulder MW, et al. Systematic review of the diagnostic accuracy of (18)F-fluorodeoxyglucose positron emission tomography in melanoma patients. *Cancer*. 2001;91(8):1530–1542.
20. Wells PS, Lensing AW, Davidson BL, et al. Accuracy of ultrasound for the diagnosis of deep venous thrombosis in asymptomatic patients after orthopedic surgery. A meta-analysis. *Ann Intern Med*. 1995;122(1):47–53.
21. Dendukuri N, Khetani K, McIsaac M, et al. Testing for HER2-positive breast cancer: a systematic review and

cost-effectiveness analysis. *CMAJ*. 2007;176(10):1429–1434.
22. Hsi ED, Tubbs RR. Guidelines for HER2 testing in the UK. *J Clin Pathol*. 2004;57(3):241–242.
23. Elkin EB, Weinstein MC, Winer EP, et al. HER-2 testing and trastuzumab therapy for metastatic breast cancer: a cost-effectiveness analysis. *J Clin Oncol*. 2004;22(5):854–863.
24. Hoang MP, Sahin AA, Ordòñez NG, et al. HER-2/neu gene amplification compared with HER-2/neu protein overexpression and interobserver reproducibility in invasive breast carcinoma. *Am J Clin Pathol*. 2000;113(6):852–859.
25. Kakar S, Puangsuvan N, Stevens JM, et al. HER-2/neu assessment in breast cancer by immunohistochemistry and fluorescence in situ hybridization: comparison of results and correlation with survival. *Mol Diagn*. 2000;5(3):199–207.
26. Bartlett JM, Going JJ, Mallon EA, et al. Evaluating HER2 amplification and overexpression in breast cancer. *J Pathol*. 2001;195(4):422–428.
27. Tsuda H, Akiyama F, Terasaki H, et al. Detection of HER-2/neu (c-erb B-2) DNA amplification in primary breast carcinoma. Interobserver reproducibility and correlation with immunohistochemical HER-2 overexpression. *Cancer*. 2001;92(12):2965–2974.
28. Press MF, Slamon DJ, Flom KJ, et al. Evaluation of HER-2/neu gene amplification and overexpression: comparison of frequently used assay methods in a molecularly characterized cohort of breast cancer specimens. *J Clin Oncol*. 2002;20(14):3095–3105.
29. Dowsett M, Bartlett J, Ellis IO, et al. Correlation between immunohistochemistry (HercepTest) and fluorescence in situ hybridization (FISH) for HER-2 in 426 breast carcinomas from 37 centres. *J Pathol*. 2003;199(4):418–423.
30. Ogura H, Akiyama F, Kasumi F, et al. Evaluation of HER-2 status in breast carcinoma by fluorescence in situ hybridization and immunohistochemistry. *Breast Cancer*. 2003;10(3):234–240.
31. Lal P, Salazar PA, Hudis CA, et al. HER-2 testing in breast cancer using immunohistochemical analysis and fluorescence in situ hybridization: a single-institution experience of 2,279 cases and comparison of dual-color and single-color scoring. *Am J Clin Pathol*. 2004;121(5):631–636.
32. Lottner C, Schwarz S, Diermeier S, et al. Simultaneous detection of HER2/neu gene amplification and protein over-expression in paraffin-embedded breast cancer. *J Pathol*. 2005;205(5):577–584.
33. Loring P, Cummins R, O'Grady A, et al. HER2 positivity in breast carcinoma: a comparison of chromogenic in situ hybridization with fluorescence in situ hybridization in tissue microarrays, with targeted evaluation of intratumoral heterogeneity by in situ hybridization. *Appl Immunohistochem Mol Morphol*. 2005;13(2):194–200.
34. Lebeau A, Deimling D, Kaltz C, et al. Her-2/neu analysis in archival tissue samples of human breast cancer: comparison of immunohistochemistry and fluorescence in situ hybridization. *J Clin Oncol*. 2001;19(2):354–363.
35. McCormick SR, Lillemoe TJ, Beneke J, et al. HER2 assessment by immunohistochemical analysis and fluorescence in situ hybridization: comparison of HercepTest and PathVysion commercial assays. *Am J Clin Pathol*. 2002;117(6):935–943.
36. Roche PC, Suman VJ, Jenkins RB, et al. Concordance between local and central laboratory HER2 testing in the breast intergroup trial N9831. *J Natl Cancer Inst*. 2002;94(11):855–857.
37. Mrozkowiak A, Olszewski WP, Piaścik A, et al. HER2 status in breast cancer determined by IHC and FISH: comparison of the results. *Pol J Pathol*. 2004;55(4):165–171.
38. Yaziji H, Goldstein LC, Barry TS, et al. HER-2 testing in breast cancer using parallel tissue-based methods. *JAMA*. 2004;291(16):1972–1977.
39. Dolan M, Snover D. Comparison of immunohistochemical and fluorescence in situ hybridization assessment of HER-2 status in routine practice. *Am J Clin Pathol*. 2005;123(5):766–770.
40. Press MF, Sauter G, Bernstein L, et al. Diagnostic evaluation of HER-2 as a molecular target: an assessment of accuracy and reproducibility of laboratory testing in large, prospective, randomized clinical trials. *Clin Cancer Res*. 2005;11(18):6598–6607.
41. Tosteson AN, Weinstein MC, Wittenberg J, et al. ROC curve regression analysis: the use of ordinal regression models for diagnostic test assessment. *Environ Health Perspect*. 1994;102(suppl 8):73–78.
42. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7(4):457–472.
43. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. London, United Kingdom: Chapman & Hall Ltd; 1995.
44. de Groot JA, Janssen KJ, Zwinderman AH, et al. Correcting for partial verification bias: a comparison of methods. *Ann Epidemiol*. 2011;21(2):139–148.
45. Lu Y, Dendukuri N, Schiller I, et al. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. *Stat Med*. 2010;29(24):2532–2543.
46. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*. 1995;141(13):263–272.