Adjusting for Differential-verification Bias in Diagnostic-accuracy Studies

A Bayesian Approach

Joris A. H. de Groot,[a] *Nandini Dendukuri,*[b] *Kristel J. M. Janssen,*[a] *Johannes B. Reitsma,*[c] *Patrick M. M. Bossuyt,*[c] *and Karel G. M. Moons*[a]

Abstract: In studies of diagnostic accuracy, the performance of an index test is assessed by verifying its results against those of a reference standard. If verification of index-test results by the preferred reference standard can be performed only in a subset of subjects, an alternative reference test could be given to the remainder. The drawback of this so-called differential-verification design is that the second reference test is often of lesser quality, or defines the target condition in a different way. Incorrectly treating results of the 2 reference standards as equivalent will lead to differential-verification bias. The Bayesian methods presented in this paper use a single model to (1) acknowledge the different nature of the 2 reference standards and (2) make simultaneous inferences about the population prevalence and the sensitivity, specificity, and predictive values of the index test with respect to both reference tests, in relation to latent disease status. We illustrate this approach using data from a study on the accuracy of the elbow extension test for diagnosis of elbow fractures in patients with elbow injury, using either radiography or follow-up as reference standards.

(*Epidemiology* 2011;22: 234–241)

In studies of diagnostic accuracy, performance of the (index) test under study is ideally determined by verifying its results against a reference standard applied to the same patients.[1] However, verification of index tests results by the preferred reference standard may not be performed in all study subjects if the standard is invasive or costly, or if a study uses retrospectively collected or routine-care data.[2,3]

Incomplete verification by the preferred reference standard can lead to bias in 2 ways.[4] The first occurs when the analysis is limited to the subset of subjects who receive the preferred reference standard. This leads to partial-verification bias, a common problem for which several solutions have been proposed.[5–8] The second occurs in studies where an alternative reference test is given to those subjects in whom the result of the preferred reference test is not available. This seems logical, but bias arises when results of the alternative reference standard are treated as if from the preferred reference standard. The reason is because the 2 reference standards are often of different quality, or they define the target condition differently.[9–11] Combining the results in a single analysis is therefore not a valid reflection of disease presence or absence as would be obtained if all subjects underwent the preferred reference standard test—thus leading to differential-verification bias.[12,13]

Figure 1A shows the analysis commonly applied in studies of a dichotomous index test in which differential disease verification is used. Results of the 2 sets of "index test-reference standard" are simply combined to achieve one "overall" table. This table is then used to estimate the accuracy of the index test in the traditional way.

For example, in a recent study of the elbow extension test by Appelboam et al,[14] all adult patients who had a positive index test would undergo radiography as the preferred reference standard, whereas patients with a negative result were verified using a structured follow-up assessment. Results of the 2 reference tests were then combined (Fig. 1B).

In the presence of differential verification, only the predictive values of the index test, with respect to each reference standard separately, are valid and interpretable using the separate 2-by-2 tables (tables (i) and (ii) in Fig. 1A). The sensitivity and specificity obtained from table (iii) in Figure 1A, are incorrect and perhaps even meaningless, as we will demonstrate.

Another widely recognized problem in diagnostic studies is that the reference standard is seldom perfect.[15,16] In most studies that use a differential-verification design, at least

Submitted 7 May 2010; accepted 14 September 2010; posted 12 January 2011.
From the [a]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands; [b]Division of Clinical Epidemiology, McGill University, Montreal, Quebec, Canada; and [c]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherlands.
Supported by The Netherlands Organization for Scientific Research (project 9120.8004 and 918.10.615).
SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).
Correspondence: Joris A. H. de Groot, Julius Center for Health Sciences and Primary care, UMC Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands. Email: J.degroot-17@umcutrecht.nl or www.juliuscenter.nl.

ISSN: 1044-3983/11/2202-0234
DOI: 10.1097/EDE.0b013e318207fc5c

A

(i) Verification with preferred reference test

	R+	R-
T+	a	b
T-	-	-

+

(ii) Verification with alternative reference test

	S+	S-
T+	-	-
T-	c	d

=

(iii) Verification with either reference test

	? +	? -
T+	a	b
T-	c	d

B

(i) Verification with radiography

	Fracture	No fracture
EET+	311	336
EET-	2 [a]	56 [a]

+

(ii) Verification with follow-up

	Fracture	No fracture
EET+	NA	NA
EET-	3	250

=

(iii) Verification with either reference test

	? +	? -
EET+	311	336
EET-	5	306

FIGURE 1. Ambiguous method to address differential verification. A, Frequently used but conceptually wrong method to handle differential verification. *T* indicates (index) text under study; *R*, preferred reference standard; *S*, alternative reference standard. B, Numerical example based on paper by Appleboam et al (adapted from *BMJ.* 2008;337:a2428). Total number of patients (adults only) who underwent radiology or follow-up as verification for their elbow extension test (EET) results. [a]Due to protocol violations a random sample of adult patients who tested negative on EET received radiography. NA indicates that results are not available and can not be estimated, because all patients who were positive on the index test underwent only the preferred reference test (radiography) and not the alternative reference method (clinical follow-up).

the alternative reference is not perfect with respect to the defined target condition. Ignoring the imperfect nature of a reference test leads to biased estimates of test accuracy due to reference standard bias.[17,18] A number of recent papers have described a Bayesian approach for correcting for partial-verification bias alone[19,20] or together with reference standard bias.[10] In addition to allowing for a more realistic model, the Bayesian approach can deal with nonidentifiability.[21]

In this study, we propose a Bayesian model to simultaneously adjust for both differential-verification bias and the imperfect nature of one or both reference standards. The model assumes that the index test, as well as both reference standards, measures a common latent variable (ie, the theoretically defined disease status).[22,23] In its most general form, the model allows for estimation of predictive values, sensitivity, and specificity of the index test, as well as both reference standards, with respect to the latent disease status. It also allows for the estimation of accuracy of the index test with respect to each reference standard. First, we will explain the model and illustrate its performance with simulated data in 2 frequently encountered scenarios of differential verification. We then illustrate its application to a real-life problem.

THE MODEL

A diagnostic study with differential verification is assumed to comprise 3 stages: Stage I, where results of the index test are collected on all study subjects; Stage II, where it is determined (by physicians or researchers) which reference test is used to verify disease status; and Stage III, where the results of the selected reference standard are collected.

Let T denote the index test. Let T_1 and T_0 be the observed number of positive and negative index test results, respectively, in the sample of $T_1 + T_0 = N$ subjects. Table 1 illustrates how the T_j subjects, $j = 0, 1$, can be subdivided into R_{j1} and R_{j0} who tested positive or negative on the preferred reference test R, and S_{j1} and S_{j0} who tested positive or negative on the alternative reference test S, so that $R_{j1} + R_{j0} + S_{j1} + S_{j0} = T_j$. Let vR_j be the probability that subjects with index test result j were verified by the preferred reference test R within group T_j, $j = 0, 1$. For simplicity, we assume the probability of verification by R or S depends only on the subject's result on the index test. Finally, let D denote the (latent) disease status taking values 1 (positive) or 0 (negative) that is the target condition for both the index test and both reference tests.

We used a Bayesian approach to estimate the unknown parameters in the model. The information from the observed data is summarized into a likelihood function. Any information on the unknown model parameters prior to data collection is summarized in terms of their joint prior probability distribution. The prior distribution is updated with the likelihood using Bayes' theorem to obtain a joint posterior distribution for the parameters. We first describe the contribution to the likelihood function by each stage of the study.

Stage I

The probabilities of testing positive or negative on the index test T are a function of the prevalence of the target condition (π), and the sensitivity (sT) and specificity (cT) of index test T:

$$P(T=1) = P(D=1)\,P(T=1|D=1) + P(D=0)\,P(T=1|D=0) = \pi sT + (1-\pi)(1-cT)$$

TABLE 1. Design of a Diagnostic Accuracy Study Using Differential Verification

Stage of Study		$T=1$	$T=0$
Stage I		T_1	T_0
Stage II	Probability of verification on R	vR_1T_1	vR_0T_0
	Probability of verification on S	$(1-vR_1)T_1$	$(1-vR_0)T_0$
Stage III	$R=1$	R_{11}	R_{01}
	$R=0$	R_{10}	R_{00}
	$S=1$	S_{11}	S_{01}
	$S=0$	S_{10}	S_{00}

T indicates (index) test; *R*, preferred reference test; *S*, alternative reference test; vR_1, probability of index test positives verified using the preferred reference test; vR_0, probability of index test negatives verified using the preferred reference test.

The likelihood contribution of the first stage is the probability of observing T1 positive results on T:

$$\propto(\pi sT + (1 - \pi)(1 - cT))^{T_1} (1 - (\pi sT + (1 - \pi)(1 - cT)))^{T_0}$$

Stage II

The contribution to the likelihood from stage II is the product of 2 independent binomial distributions corresponding to the probability of verification by the preferred reference standard within the 2 groups $T = 1$ and $T = 0$:

$$\propto vR_1^{(R_{11} + R_{10})}(1 - vR_1)^{(S_{11} + S_{10})} vR_0^{(R_{01} + R_{00})} (1 - vR_0)^{(S_{01} + S_{00})}$$

Stage III

In Stage III, we estimate the predictive values of the index test with respect to each reference standard. We assume T is conditionally independent of both reference tests, given the true disease status. The predictive values can then be expressed as functions of the prevalence and the sensitivity and specificity of the index and reference tests.

For reference standard R, we have,[10]

$$P(R = 1|T = 1) = sR \frac{(\pi sT)}{(\pi sT + (1 - \pi)(1 - cT))} + (1 - cR) \frac{((1 - \pi)(1 - cT))}{(\pi sT + (1 - \pi)(1 - cT))}$$

Similarly,

$$P(R = 1|T = 0) = sR \frac{(\pi(1 - sT))}{(\pi(1 - sT) + (1 - \pi)cT)} + (1 - cR) \frac{((1 - \pi)cT)}{(\pi(1 - sT) + (1 - \pi)cT)}$$

where sR and cR are the sensitivity and specificity of the preferred reference standard with respect to the latent true disease status, D. In the particular case when the preferred reference test is considered perfect (ie, $sR = cR = 1$), these expressions reduce to:

$$P(R = 1|T = 1) = \frac{(\pi sT)}{(\pi sT + (1 - \pi)(1 - cT))} \text{ and}$$

$$P(R = 1|T = 0) = \frac{(\pi(1 - sT))}{(\pi(1 - sT) + (1 - \pi)cT)}$$

We can derive similar expressions for reference standard S.

The contribution to the likelihood of this stage is the product of 4 independent binomial density functions, each corresponding to the probability of a positive result on a reference test conditional on the index test:

$$\propto P(R = 1|T = 1)^{R_{11}}(1 - P(R = 1|T = 1))^{R_{10}} P(R = 1|T = 0)^{R_{01}}(1 - P(R = 1|T = 0))^{R_{00}}.$$

$$P(S = 1|T = 1)^{S_{11}}(1 - P(S = 1|T = 1))^{S_{10}} P(S = 1|T = 0)^{S_{01}}(1 - P(S = 1|T = 0))^{S_{00}}$$

Model Identifiability

There are 9 unknown parameters—sensitivities and specificities of each of the 3 tests, prevalence and verification probabilities. However, there are only 7 degrees of freedom—1 from Stage I, 2 from Stage II, and 4 from Stage III. To solve this nonidentifiable[10,21] problem, we need to provide informative prior distributions for at least $9 - 7 = 2$ of the parameters. To be precise, we need to provide informative prior distributions on any 2 parameters involved in Stages I and III (π, sT, cT, sR, or cR). The 2 verification probabilities do not affect these parameters and may be estimated with low-information prior distributions. In the case when the preferred reference standard is considered perfect ($sR = cR = 1$), the number of unknown parameters is 7 and the model is identifiable.

Prior Distributions

Following others, we used independent Beta (α,β) prior distributions for each unknown parameter because they cover the [0,1] range and have a flexible shape, making it easy to match the density to prior information.[21] To determine the values of α and β for a parameter about which substantive prior knowledge is available (eg, sensitivity or specificity of the preferred reference test), we need information on any 2 features of the distribution, eg, mean and standard deviation.[21] For both simulations and the real-life application, we elicited prior information on sensitivity and specificity parameters in the form of a range of plausible values. Parameters of the corresponding Beta prior distributions were determined by assuming that the middle point of the range was equal to the prior mean (μ) and one-quarter of the range was equal to the prior standard deviation (σ).

Based on these assumptions, we determined α and β as follows:

$$\alpha = -\frac{\mu(\sigma^2 + \mu^2 - \mu)}{\sigma^2} \text{ and } \beta = \frac{(\mu - 1)(\sigma^2 + \mu^2 - \mu)}{\sigma^2}$$

It is realistic to assume that there is prior information available about the sensitivity and specificity of the reference tests R and S. The fact that they are used as reference tests indicates that their properties with regard to the target con-

dition are at least approximately known from past experience, or by comparison with more accurate disease-detection methods (such as autopsy). Regarding the index test, T, one typically would want to apply low-information prior distributions on sT and cT, so as to limit the incorporation of any subjective prior opinions about the main parameters of interest. Therefore, we used the uniform Beta ($\alpha = 1$, $\beta = 1$) distribution.

Estimation of the Posterior Distribution

The data are combined through the likelihood function with the prior distribution to derive posterior distributions using Bayes' theorem.[21] We base our posterior inferences on samples from the joint posterior distribution obtained using the WinBUGS software program (eAppendix, http://links.lww.com/EDE/A446). For each application in this paper, we ran 5 chains with different starting values. Each chain had a total of 20,000 iterations, of which we dropped the first 2000 to allow for a burn-in period. Convergence of the Markov Chain Monte Carlo sampling was checked using the Gelman-Rubin statistic.[24] Summary statistics (median, 2.5% and 97.5% quantiles) of parameters of interest were then estimated.

TWO SIMULATED EXAMPLES

We use simulated data to describe 2 prototypical differential-verification designs. The 2 designs differ in terms of disease verification strategy. Parameter values used to simulate the scenarios are summarized in Table 2. In design 1, all subjects who test positive on the index test are verified with preferred reference test R, while all who test negative are—usually by design—verified with reference test S (Table 3, top).

In design 2, a large proportion of the index test positives (70%) and a small proportion of the test negatives (30%) will be verified by the preferred reference test R (Table 3, bottom), while the remainder are verified with S. This design is often encountered when disease verification is less strictly designed, ie, when performing the preferred or second reference test is not according to a protocol.[2,13]

The purpose of these simulations is to illustrate the bias that may arise when ignoring differential verification, and also to show that our method will generally result in posterior credible intervals that capture the true value of the parameters when the informative prior distributions are correctly specified.

TABLE 2. True Parameter Values for the 2 Simulated Designs With Different Strategies of Differential Verification

Design 1			Design 2		
	Sensitivity With Respect to D[a]	Specificity With Respect to D[a]		Sensitivity With Respect to D[a]	Specificity With Respect to D[a]
T	0.7	0.7	T	0.7	0.7
R	0.95	0.95	R	0.95	0.95
S	0.85	0.85	S	0.85	0.85
π	0.2		π	0.2	
vR_1	1		vR_1	0.7	
vR_0	0		vR_0	0.3	

[a]True sensitivities and specificities with respect to the disease status (D). D indicates true disease status.

TABLE 3. Simulated Data Sets From 2 Differential-verification Designs

Stage of Study		$T = 1$	$T = 0$
Data From Differential-verification Design 1			
Stage I		304	496
Stage II	Probability of verification on R	1	0
	Probability of verification on S	0	1
Stage III	$R=1$	116	0
	$R=0$	188	0
	$S=1$	0	108
	$S=0$	0	388
Data From Differential-verification Design 2			
Stage I		304	496
Stage II	Probability of verification on R	0.7	0.3
	Probability of verification on S	0.3	0.7
Stage III	$R=1$	81	20
	$R=0$	132	128
	$S=1$	37	76
	$S=0$	54	272

Analysis of the 2 Simulated Data Sets

We compare the results from the analysis of these data with our model to results from separate cross-tabulations of T versus R and T versus S simply added together, as described in Figure 1.

We used Beta (71.25, 3.75) prior distributions for both sR and cR, corresponding to a density centered at 0.95 and a range of 0.90–1.00. For the sensitivity and specificity of S, we used Beta (172.55, 30.45), corresponding to a density centered at 0.85 and a range of 0.80–0.90.[21] We used low information priors (Beta (1,1)) for sT, cT π, vR_1, and vR_0.

Results From Differential-verification Design 1

Results for differential-verification design 1 are summarized in Table 4. The Bayesian model provided median estimates close to the true values for all accuracy measures of the index test (Table 2). In addition to the results in Table 4, the model also provides estimates for sR (0.95 [95% credible interval (CI) = 0.89–0.99]) and cR (0.95 [0.89–0.99]) of the preferred reference test R with respect to the true disease status, as well as the sensitivity (0.85 [0.80–0.90]) and specificity (0.85 [0.80–0.89]) of the second reference test S.

TABLE 4. Summary of Results[a] of Analyses of Simulated Data From Differential-verification Design 1

	Sensitivity (CI)	Specificity (CI)	PPV (CI)	NPV (CI)
Accuracy measures *T* with respect to *D*				
Truth	0.70	0.70	0.37	0.90
Bayesian approach	0.73 (0.57–0.94)	0.70 (0.66–0.74)	0.37 (0.30–0.43)	0.91 (0.83–0.98)
Accuracy measures *T* with respect to *R*				
Truth	0.63	0.69	0.38	0.86
Bayesian approach	0.64 (0.50–0.82)	0.7 (0.65–0.74)	0.38 (0.33–0.43)	0.87 (0.79–0.95)
Analysis (separate tables)	NA	NA	0.38 (0.33–0.44)	NA
Accuracy measures *T* with respect to *S*				
Truth	0.53	0.68	0.41	0.78
Bayesian approach	0.54 (0.48–0.60)	0.68 (0.64–0.72)	0.41 (0.35–0.47)	0.78 (0.75–0.82)
Analysis (separate tables)	NA	NA	NA	0.78 (0.74–0.82)
Accuracy measures *T* with respect to ?				
Combining 2 tables	0.52 (0.45–0.58)	0.67 (0.63–0.71)	0.38 (0.33–0.44)	0.78 (0.74–0.82)

[a]Posterior median and 95% credible intervals.
PPV indicates positive predictive value; NPV, negative predictive value; CI, credible interval.

The estimated prevalence of the (latent) disease status was correctly estimated at 0.20 ([0.14–0.25]).

For the naive analysis, the 116 + 108 subjects who tested positive on either of the reference tests are considered true positives. Similarly, the 188 + 388 subjects who were negative on either reference test are considered true negatives. The resulting estimates of sensitivity and specificity and predictive values of the index test are thus measured with respect to neither *R* nor *S*. Neither the true value of the sensitivity of *T* with respect to *D* nor with respect to *R* is captured within the credible interval for the sensitivity of *T*, based on the naive analysis (0.52 [0.45–0.58]). For this particular design, only the positive predictive value with respect to *R* and the negative predictive value with respect to *S* can be obtained without bias from the combined table.

The small number of patients verified in this example results in a considerably smaller Stage III sample compared with the Stage I sample. Therefore, the Bayesian estimates of sensitivity and specificity become less precise. These wide(r) credible intervals suggest that this type of design would require a large sample size to obtain a meaningful precision.

Results From Differential-verification Design 2

Results for differential-verification design 2 are summarized in Table 5. The Bayesian model again provides good estimates for all accuracy measures. The model also provides

TABLE 5. Summary of Results[a] of Analyses Based on Simulated Data From Differential-verification Design 2

	Sensitivity (CI)	Specificity (CI)	PPV (CI)	NPV (CI)
Accuracy measures *T* with respect to *D*				
Truth	0.70	0.70	0.37	0.90
Bayesian approach	0.71 (0.58–0.86)	0.70 (0.66–0.74)	0.37 (0.30–0.45)	0.90 (0.84–0.97)
Accuracy measures *T* with respect to *R*				
Truth	0.63	0.69	0.38	0.86
Bayesian approach	0.63 (0.54–0.73)	0.70 (0.66–0.73)	0.38 (0.32–0.44)	0.86 (0.82–0.91)
Analysis (separate tables)	0.80 (0.71–0.87)	0.49 (0.43–0.56)	0.38 (0.32–0.45)	0.86 (0.80–0.91)
Accuracy measures *T* with respect to *S*				
Truth	0.53	0.68	0.41	0.78
Bayesian approach	0.53 (0.47–0.59)	0.68 (0.65–0.72)	0.41 (0.36–0.46)	0.78 (0.74–0.82)
Analysis (separate tables)	0.33 (0.24–0.42)	0.83 (0.79–0.87)	0.41 (0.31–0.52)	0.78 (0.74–0.82)
Accuracy measures *T* with respect to ?				
Combining 2 tables	0.55 (0.48–0.62)	0.68 (0.64–0.72)	0.39 (0.33–0.45)	0.81 (0.77–0.84)

[a]Posterior median and 95% credible intervals.

estimates for *sR* (0.95 [0.89–0.99]) and *cR* (0.95 [0.90–0.99]), the sensitivity (0.85 [0.80–0.90]) and specificity (0.85 [0.81–0.89]) of the second reference test *S* with respect to the true disease status, and π (0.20 [0.14–0.25]). We carried out a naive analysis comparing *T* to the result of either reference standard combined, as well as a separate analysis of the tables comparing *T* to *R* or *S*. Once again, the sensitivity estimate from the naive analysis (0.55 [0.48–0.62]) is biased, with its 95% credible interval capturing neither the true value of *sT* nor *sR*. Notice that when combining the tables under design 2, even the predictive values with respect to *R* and *S* become biased. However, the predictive values of *T* with respect to *R* and *S* are both appropriately estimated when using the data in separate tables of *T* versus *R* and *T* versus *S*.

APPLICATION TO A REAL-LIFE PROBLEM

In the recent study by Appelboam et al[14] on the elbow extension test to rule out elbow fracture in adults (and children), a differential-verification design was used. Their preferred reference test to verify whether patients had an elbow fracture was radiography. For unstated reasons (most likely costs or radiation reduction), they planned to perform radiography only in patients with a positive elbow extension test.

However, due to protocol violations, a small subset of patients with a negative elbow extension test also received radiography. The protocol violations occurred mostly when temporary staff misunderstood or were unaware of the protocol, suggesting that this was most likely a random subset of negative-test patients. All remaining negative-test patients who did not undergo radiography received a structured follow-up assessment (the alternative reference test) by telephone to verify whether indeed elbow fracture was absent. Patients who met any of the prespecified recall criteria were asked to return for radiography. Those not requiring recall were assumed not to have an elbow fracture. The resulting data (for adults) were shown in Figure 1B.

In consultation with experts in orthopedics and radiology, and after reviewing the literature,[25,26] we determined the range of sensitivity and specificity for both reference tests with respect to the defined target condition (ie, all elbow fractures) (Table 6). Radiography is believed to have both high sensitivity and high specificity, while follow-up is believed to have slightly better sensitivity but much worse specificity.

Although the opinions of experts are unlikely to be polarized with regard to the accuracy of the 2 reference tests, there may be a debate on the form of the prior distribution. We therefore carried out a sensitivity analysis using a uniform prior distribution over the same range of values to see to what degree the priors affect our results. As it happens, the results did not change greatly.

Appelboam et al[14] reported overall estimates of accuracy of the elbow extension test, ignoring the use of different reference standards. These were interpreted as estimates of accuracy with respect to radiography. Though both radiography and structured follow-up are useful verification methods, their results are not interchangeable, as discussed earlier in the text. We assume both are imperfect measures of the latent variable "all elbow fractures."

TABLE 6. Plausible Ranges for Sensitivity and Specificity of the Reference Tests and Corresponding Coefficients of the Beta Prior Densities Used to Analyze the Data From the Elbow Fracture Study

	Radiography			Structured Follow-up		
		Beta Coefficients			Beta Coefficients	
	Range (%)	α	β	Range (%)	α	β
Sensitivity	90–100	71.3	3.8	95–100	151.1	3.9
Specificity	95–100	151.1	3.9	40–60	49.5	49.5

We used our model to estimate accuracy of the elbow extension test with respect to radiography and structured follow-up separately. Because this is a typical situation where the proportions of verified subjects (vR_1 and vR_0) are not predetermined fixed numbers, we did not use fixed numbers; instead, distributions as defined in the formula in the model section under Stage II. By adjusting for the imperfect sensitivity and specificity of the reference standards, we also estimate the accuracy of the index test with respect to the latent target condition. Informative Beta prior distributions over the sensitivity and specificity of the reference tests were determined using the ranges in Table 6. The results of the Bayesian estimation in the form of posterior medians and 95% credible intervals appear in Table 7.

We can see that the accuracy measures differed greatly with respect to each reference standard. This is due partly to the fact that the population who undergoes the preferred reference standard will be at higher risk for having the disease than the population who undergoes the alternative reference standard, because selection is based on one or more diagnostic test results. This underlines the importance of reporting clearly the theoretical or clinical reference standard on which the accuracy of an index test is based.

As is commonly done, Appelboam et al[14] interpreted the combination of the 2 reference standards as 1 gold standard (Fig. 1) and calculated the accuracy measures of the elbow extension test accordingly. This resulted in relatively higher sensitivity (98; 95% CI = 96–100) and negative predictive value (98 [96–100]), as compared with those based on radiography (sensitivity = 97 [95–99]; negative predictive value = 97 [94–99]) and follow-up (88 [86–91] and 85 [81–88]). These adjusted measures of the reference test are, in our view, of more clinical relevance than the overall measures.

TABLE 7. Accuracy Measures of the Elbow Extension Test to Diagnose Elbow Fracture (in Adults)

Method	Sensitivity (CI)	Specificity (CI)	NPV (CI)	PPV (CI)
Appleboam et al[14]				
Uncorrected accuracy with respect to combined reference test	98.4 (96.3–99.5)	47.7 (43.7–51.6)	98.4 (96.3–99.5)	48.1 (44.2–52.0)
Using the Bayesian model				
Corrected accuracy with respect to (latent) target condition	99.6 (98.4–100.0)	48.6 (44.5–52.9)	99.5 (98.4–100.0)	49.2 (44.3–54.6)
Corrected accuracy with respect to radiography	97.0 (94.5–98.8)	47.4 (43.5–51.3)	96.9 (94.3–98.8)	48.0 (44.2–51.9)
Corrected accuracy with respect to follow-up	88.3 (85.5–90.9)	47.9 (43.8–52.2)	84.7 (80.8–88.3)	55.5 (50.8–60.5)

CI indicates credible interval; NPV, negative predictive value; PPV, positive predictive value.

The model also produced accuracy values with respect to the (latent) true disease status. In this example, the elbow extension test had extremely high sensitivity (99.7 [98.4–100]) and negative predictive value (99.8 [98.7–100]) with respect to the defined "latent" disease status of all elbow fractures.

In addition, the model provided estimates for the sensitivity (95 [89–99]) and specificity (97 [95–99]) of radiography (with respect to the [latent] target condition), as well as the sensitivity (98 [95–99]) and specificity (87 [84–90]) of the follow-up. The estimated prevalence was 30% (95% CI = 27–34).

DISCUSSION

We have presented a Bayesian approach for simultaneously adjusting for differential-verification bias and multiple imperfect reference standards, in diagnostic studies aimed at estimating the predictive values, sensitivity, and specificity of a single index test.

The model produces accuracy measures with respect to both the (latent) disease status and the separate reference standards. The former can be considered as a more general measure of performance of the index test with respect to a theoretically defined disease status, in case none of the reference standards is "perfect." The index tests' accuracy measures per reference standard, may, however, be considered of greater clinical relevance. These measures reflect the accuracy against the reference tests that are performed in clinical practice, on which further patient management decisions will be based.

Related to this, various reference standards commonly differ in their definition of the target condition. For example, in patients suspected of appendicitis, one may have data on histopathology of the appendix or on clinical follow-up. Histopathology seems to be the preferred reference test because it reveals even the smallest amount of inflamed cells. However, in clinical practice, the more interesting information is not whether the patient has inflamed cells, but whether the patient recovers without intervention. This would make follow-up the clinically preferred reference. Even though it would not be feasible (and indeed would be unethical) to rely on follow-up in every patient, this example shows that different reference tests can address slightly different target conditions. In this case, and more generally in the absence of a single reference standard (eg, for testing heart failure, Alzheimer disease, or diabetes), a precise definition of the disease or latent (disease) class is of utmost importance.

We made some simplifying assumptions to facilitate a clear presentation. We assumed that the probability of verification by reference test R or S depends only on the results of the index test. We are aware that in clinical practice a test is always judged in the context of other information.[27,28] Our method, however, can be extended to study the accuracy of a multivariable model to allow the probability of verification by R or S to depend on more information than T alone. Another assumption that may be questioned is the conditional independence between the index test and each reference standard. Once again, we are aware that this may affect our estimates.[29,30] The models, however, can be extended to incorporate conditional dependence. Proper discussion of these 2 extensions would merit a separate article with extensive simulations. Such additional simulations would also give more insight into what factors (eg, prevalence and correlation between index test and alternative reference test) have influence on the direction and the magnitude of the differential-verification bias.

The important message is to avoid verification bias in diagnostic studies by verifying as many patients as possible with the preferred reference standard. Complete verification may not be possible for various reasons, such as patient burden and costs.[2] In situations where verification by the preferred reference standard is impossible or unethical in specific groups, verification by a different reference standard can be considered. Overall accuracy estimates that ignore the use of different reference standards are difficult to interpret, and results should be reported separately for each reference standard to provide informative and unbiased measures of accuracy. To evaluate the index test with regard to the true target condition of interest, one should also correct for possible imperfection of the reference standards used. The method we present may help researchers make unbiased

inferences about a variety of index test characteristics in the presence of differential verification.

REFERENCES

1. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, ed. *The Evidence Base of Clinical Diagnosis.* 2nd ed. London: BMJ Books; 2002:39–60.
2. Oostenbrink R, Moons KG, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. *J Clin Epidemiol*. 2003;56:501–506.
3. van der Schouw YT, Van DR, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. *J Clin Epidemiol*. 1995;48:417–422.
4. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J Clin Epidemiol*. 2009;62:797–806.
5. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39:207–215.
6. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. *Stat Methods Med Res*. 1998;7:337–353.
7. Harel O, Zhou XH. Multiple imputation for correcting verification bias. *Stat Med*. 2006;25:3769–3786.
8. de Groot JA, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB. Multiple imputation to correct for partial verification bias revisited. *Stat Med*. 2008;27:5880–5889.
9. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004;140:189–202.
10. Lu Y, Dendukuri N, Schiller I, Joseph L. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. *Stat Med*. 2010;29:2532–2543.
11. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061–1066.
12. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess.* 2007;11(50):1–51.
13. Rutjes AW, Reitsma JB, Irwig L, Bossuyt PM. Partial and differential verification in diagnostic accuracy studies. Sources of bias and variation in diagnostic accuracy studies [dissertation]. Amsterdam: Academic Medical Center; 2005:31–44. Available at http://dare.uva.nl/record/168357.
14. Appelboam A, Reuben AD, Benger JR, et al. Elbow extension test to rule out elbow fracture: multicentre, prospective validation and observational study of diagnostic accuracy in adults and children. *BMJ*. 2008;337:a2428.
15. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6:411–423.
16. Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. *J Chronic Dis*. 1981;34:599–610.
17. Plassman BL, Khachaturiane AS, Townsend JJ, et al. Comparison of clinical and neuropathologic diseases of alzheimers disease in 3 epidemiologic samples. *Alzheimers Dement*. 2006;2:2–11.
18. Wiederkehr S, Simard M, Fortin C, van RR. Validity of the clinical diagnostic criteria for vascular dementia: a critical review. Part II. *J Neuropsychiatry Clin Neurosci*. 2008;20:162–177.
19. Buzoianu M, Kadane JB. Adjusting for verification bias in diagnostic test evaluation: a Bayesian approach. *Stat Med*. 2008;27:2453–2473.
20. Martinez EZ, Achcar JA, Louzada-Neto F. Estimators of sensitivity and specificity in the presence of verification bias: a Bayesian approach. *Comput Stat Data Anal*. 2006;51:601–611.
21. 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;141:263–272.
22. Kaldor J, Clayton D. Latent class analysis in chronic disease epidemiology. *Stat Med*. 1985;4:327–335.
23. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J Clin Epidemiol*. 1988;41:923–937.
24. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7:457–511.
25. McGinley JC, Roach N, Hopgood BC, Kozin SH. Nondisplaced elbow fractures: a commonly occurring and difficult diagnosis. *Am J Emerg Med*. 2006;24:560–566.
26. Pudas T, Hurme T, Mattila K, Svedstrom E. Magnetic resonance imaging in pediatric elbow fractures. *Acta Radiol*. 2005;46:636–644.
27. Moons KG, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic studies? *J Clin Epidemiol*. 2002;55:633–636.
28. Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. *Clin Chem*. 2004;50:473–476.
29. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*. 2001;57:158–167.
30. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*. 1985;41:959–968.