Diagnostic research in the presence of an incomplete or imperfect reference standard

J.A.H. de Groot

Diagnostic research in the presence of an incomplete or imperfect reference standard.

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht.
Thesis. University Utrecht, Faculty of medicine, with a summary in Dutch.

ISBN: 978-94-6108-198-8
Cover Design: www.theupperroom.nl
Print: Drukkerij Gildeprint, Enschede

Diagnostic research in the presence of an incomplete or imperfect reference standard

Diagnostisch onderzoek met een incomplete of imperfecte referentie standaard

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 13 september 2011 des middags te 4.15 uur

door

Joris Antoon Harry de Groot
geboren op 30 Juli 1980, te Rosmalen

Promotoren: Prof.dr. K.G.M. Moons

Prof.dr. P.M. Bossuyt

Co-promotoren: Dr. K.J.M. Janssen

Dr. J.B. Reitsma

The studies in this thesis were funded by the Netherlands Scientific Organization (ZonMw 912-08-004 en ZonMW 918-10-615).

Financial support by the Julius Center for Health Sciences and Primary Care for the publication of this thesis is gratefully acknowledged.

Manuscripts based on the studies presented in this thesis

Chapter 2
de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, Moons KG. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ 2011;343:d4770

Chapter 3.1
de Groot JA, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB. Multiple Imputation to correct for partial verification bias revisited. Stat Med. 2008; 27(28):5880–5889

Chapter 3.2
de Groot JA, Janssen KJ, Zwinderman AH, Bossuyt PM, Reitsma JB, Moons KG. Correcting for Partial Verification Bias: A Comparison of Methods. Ann Epidemiol. 2011; 21:139–148

Chapter 4.1
de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Bossuyt PM, Moons KG. Adjusting for differential verification bias in diagnostic-accuracy studies: A Bayesian approach. Epidemiology. 2011;22: 234–241

Chapter 5
de Groot JA, Dendukuri N, Reitsma JB, Bossuyt PM, Janssen KJ, Moons KG. Diagnostic accuracy studies without a single, acceptable reference standard: a case study.
In preparation

Chapter 6
de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Brophy J, Joseph L, Bossuyt PM, Moons KG. Adjusting for partial verification or work-up bias in meta-analyses of diagnostic accuracy studies.
Provisionally accepted by AJE

Chapter 7
de Groot JA, Moons KG., Janssen KJ, Bossuyt PM, Reitsma JB Reference standard problems in studies of diagnostic accuracy: an overview.
Submitted

Contents

Chapter 1

General introduction

Diagnostic accuracy research is a vital step in the evaluation of new diagnostic technologies.[1;2] It is the ability of a test to correctly discriminate between patients that have and do not have the target disease (or target condition). In studies of diagnostic accuracy, results of the tests under study (index tests) are compared with results of a reference standard (or reference test) applied to the same series of patients.[3] In this research framework, the reference standard is the best available method to verify the presence or absence of the target disease, and thus provides the definitive classification of patients into target disease present or absent. This process is known as disease verification. Measures such as predictive values or post-test probabilities, ROC curves, sensitivity, specificity, likelihood ratios, or odds ratio, express how well the results of the index tests (either in isolation or in combination using, e.g., a multivariable diagnostic model) agree with the outcomes or results of the reference standard.[2]

Ideally, the reference standard provides error-free disease classification. In some situations, it is not possible to verify the disease outcome with the preferred reference standard in all patients. In other situations, the reference standard cannot be applied to any patient. Failure to apply the reference standard may result in various types of disease verification problems. Biased and exaggerated estimates of accuracy of a test can lead to inefficiencies in testing in clinical practice, unnecessary costs, and could trigger physicians to making incorrect treatment decisions.

The first major problem in diagnostic research is when partial disease verification is present. This problem occurs when only a sub sample of the patients who have had the index tests subsequently receives verification by the reference standard. The disease outcome is then verified or observed in only part of the total study group. If this partial disease verification by the reference standard is completely ad random (i.e. non-selective), this process obviously poses no validity problems; it only may reduce the statistical power of the study. If this partial verification, however, is based on index test results or other observed patient information or characteristics, which is often the case in diagnostic research, the verified patients are obviously not a random subsample of the total study group. We then speak of selective disease verification, which yields selectively missing disease outcomes. If in this case the non-verified patients are simply left out of the analysis, the most common approach, the estimated measures of diagnostic accuracy of the index tests are usually biased; this bias is called partial verification bias.[4;5]

Another problem in diagnostic accuracy studies occurs when an alternative, second best, reference standard is used in those subjects where the result of the first, preferred reference test is not obtained. This second reference standard is often applied in those with

negative results on the index tests or with mild clinical presentation. Although this seems a logical and clinically appealing and ethical approach, bias can arise when the results of both reference tests are treated as interchangeable. Yet the two reference tests are almost by definition of different quality, in terms of target disease classification, or may even define the target disease differently.[6;7] Simply combining all data in a single analysis, as if both reference tests yield the same disease outcomes, does not validly reflect the 'true' disease presence/absence status; the so-estimated disease prevalence differs from what one would have obtained if all subjects had undergone the preferred reference standard.

Consequently, estimated accuracy measures of the index tests will again be biased. This bias is called differential verification bias.[8;9]

There are also situations or diseases where there an appropriate reference standard does not exist. In that case one may consider the use of an expert panel.[2;10] A panel of experts then decides on the 'true' presence or absence of the target condition in each patient, often based on all relevant or documented information of that patient. An alternative for this expert panel method is the use of statistical models for combining multiple test results to classify patients, known as latent class models.[11-13] They relate the observed patterns of various test results to unknown or latent categories which are defined by the presence or absence of the target condition. By linking index test results to these latent disease categories, such models can estimate the accuracy of the index tests. Latent class models vary in their underlying assumptions and in the way they estimate these parameters.[11;14;15] Verification problems are not only a problem in primary studies of diagnostic accuracy but also pose a challenge for systematic reviews of diagnostic studies. Some authors have acknowledged this bias in their discussions, but did not quantify or correct for it in the analyses.[16-18] So far, correction methods have not been developed and incorporated into meta-analytic statistical models when verification problems are present in one or more of the primary diagnostic studies.

Objectives

The main aim of the work reported in this thesis was to examine the problem of verification bias in studies of diagnostic accuracy. In particular, we aimed to investigate the available methods to alleviate the various problems of verification bias and, more importantly, to improve the methodology and analysis of primary diagnostic accuracy studies and diagnostic meta-analyses in the presence of various forms of verification bias.

Outline of this thesis

In **Chapter 2** we describe the most important types of disease verification problems using examples from clinical practice.

Chapter 3 contains two studies on the correction of partial verification bias in diagnostic accuracy studies. In **Chapter 3.1** we revisit and compare various methods to correct for partial verification bias. In **Chapter 3.2** we focus on the ability of multiple imputation and the correction method of 'Begg and Greenes' under a range of different situations of selectively missing disease outcomes or partial verification. We elucidate under which circumstances both methods produce similar results and when they may start to deviate. Based on our findings we propose guidance for researchers designing and analyzing diagnostic accuracy studies with partial -both selective and non-selective- disease verification.

In **Chapter 4** we discuss differential verification bias. We propose a Bayesian model to simultaneously adjust for both differential verification bias and the imperfect nature of one or both applied reference tests. In its most general form, the model allows for estimation of predictive values, sensitivity, and specificity of the index tests, as well as of both (imperfect) reference tests, with respect to the estimated latent disease status. It also allows for the estimation of accuracy of the index tests with respect to each reference tests.

In **Chapter 5** we compare the results of an expert panel and a latent class model in diagnostic accuracy studies in which there is no (single) reference method available. Using data from an empirical study on the diagnosis of heart failure we identify and discuss differences between the expert panel decision and the results of a latent class model.

In **Chapter 6** we propose a new two-stage Bayesian approach to correct for partial 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 primary studies with complete verification 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 predictive values of the index test(s). The results from the two stages can then be combined to obtain unbiased summary estimates of the sensitivity and specificity of the index test(s).

Chapter 7 provides a general discussion of the different types of disease verification problems that may occur in diagnostic research practice and concludes with various recommendations depending on the situation.

Reference List

1. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002; 324(7336):539-541.
2. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, editor. The Evidence Base of Clinical Diagnosis. 2nd ed. London: BMJ Books; 2002. 39-60.
3. Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic procedures. The evidence base of clinical diagnosis. London: BMJ Books; 2002. 1-18.
4. Reitsma JB, Rutjes AWS, Khan KS, Coomarasamy A, Bossuyt P. NHS Health Technology Assessment Programme: A systematic review of methods to evaluate tests when there is no gold standard. 2007. 6-4-2009.
5. 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(8):797-806.
6. 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(3):189-202.
7. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11):1061-1066.
8. 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 (Thesis) [Amsterdam: 2005.
9. 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):iii, ix-51.
10. Gagnon R, Charlin B, Coletti M, Sauve E, Van d, V. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? Med Educ 2005; 39(3):284-291.
11. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. Biostatistics 2007; 8(2):474-484.
12. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol 1988; 41(9):923-937.
13. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics 1996; 52(3):797-810.
14. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 2001; 57(1):158-167.

15. 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(3):263-272.
16. Met R, Bipat S, Legemate DA, Reekers JA, Koelemay MJ. Diagnostic performance of computed tomography angiography in peripheral arterial disease: a systematic review and meta-analysis. JAMA 2009; 301(4):415-424.
17. Nayak S, Olkin I, Liu H, Grabe M, Gould MK, Allen IE et al. Meta-analysis: accuracy of quantitative ultrasound for identifying patients with osteoporosis. Ann Intern Med 2006; 144(11):832-841.
18. Mijnhout GS, Hoekstra OS, van Tulder MW, Teule GJ, Deville WL. Systematic review of the diagnostic accuracy of (18)F-fluorodeoxyglucose positron emission tomography in melanoma patients. Cancer 2001; 91(8):1530-1542.

Chapter 2

Verification problems in diagnostic accuracy studies

Joris A.H. de Groot
Patrick M.M. Bossuyt
Johannes B. Reitsma
Anne W.S. Rutjes
Nandini Dendukuri
Kristel J.M. Janssen
Karel G.M. Moons

BMJ 2011;343:d4770

Abstract

Background

Diagnostic accuracy is the ability of a test or combination of tests (e.g. in a diagnostic model) to correctly identify patients with or without the target disease. In diagnostic accuracy studies, ideally all patients that undergo the index test are verified by the reference standard. Incomplete or improper disease verification is one of the major sources of bias in diagnostic accuracy studies. This study describes the various types of disease verification problems using empirical examples and proposes solutions to alleviate the associated biases.

Partial verification bias

Partial verification bias occurs when not all patients are verified by the reference standard, and when this (referral for) disease verification is related to other, previous (index) test results or patient characteristics. If the preferred reference standard has not been applied in all patients, selectively or non-selectively, mathematical correction methods can be used to correct for the partial verification bias.

Differential verification bias

Another approach in diagnostic accuracy studies is to use an alternative reference test in those subjects where the result of the first, preferred reference test cannot be obtained. Although this seems a clinically appealing and ethical approach, differential verification bias arises when the results of both reference tests are treated as equal and interchangeable when, in fact, they are of different quality or define the target condition differently. Then, the estimated accuracy of the diagnostic index test or model should be corrected and reported separately for each reference test, to provide more informative and less biased index tests' accuracy measures.

Conclusion

In diagnostic accuracy studies, efforts should be made to verify as many patients as possible – preferably all - with the optimal reference standard, to avoid bias in the estimated accuracy of the index test. Often, patient burden, costs or other reasons prevent this from happening. If so, researchers and practicing physicians should be aware of the associated biases, and provide the reader with corrected and more informative estimates of the accuracy of the diagnostic test or model under study.

Introduction

The accuracy of a diagnostic test or combination of tests (e.g. in a diagnostic model) is the ability to correctly identify patients with or without the target disease. In studies of diagnostic accuracy, results of the test or model under study are verified by comparing them with results of a so-called reference standard, applied to the same patients, to verify disease status (Figure 1A).[1] Measures such as predictive values, post-test probabilities, ROC curves, sensitivity, specificity, likelihood ratios, or odds ratios, express how well the results of an index test agree with the outcome of the reference standard.[2] Biased and exaggerated estimates of diagnostic accuracy can lead to inefficiencies in diagnostic testing in practice, unnecessary costs, and could trigger physicians to making incorrect treatment decisions.

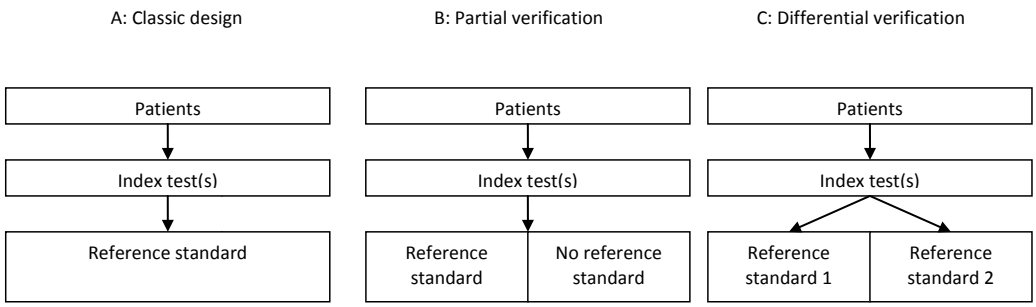

Figure 1. Diagnostic accuracy study with complete verification by the same reference standard (A); study with partial verification (B); and study with differential verification (C).

The reference standard ideally provides error-free classification of the disease outcome presence or absence. In some cases, it is not possible to verify the definitive disease presence/absence in all patients with the (single) reference standard, which may result in bias. In this paper, we describe the most important types of disease verification problems using examples from published diagnostic accuracy studies. We also propose solutions to alleviate the associated biases.

Table 1 provides a small sample of published diagnostic accuracy studies, including from our own group, in which disease verification problems were present.

Table 1. Examples of published diagnostic accuracy studies in which disease verification problems, partial or differential, were present.

Partial verification				
Index test(s)	Target condition	Reference standard	Problem	References
FDG-PET scan	Distant metastases	Histology of biopsy	Only PET hotspots were (can be) biopsied.	(3)
Digital rectal examination and prostate specific antigen	Prostate cancer	Combination of transrectal ultrasound plus biopsy	54 men did not undergo the reference standard (test combination) for unknown reasons.	(4)
Dobutamine-atropine stress echocardiography	Coronary artery disease	Coronary angiography	Only a small sample of patients received the reference because of the practitioners` decision.	(5)
Hepatic scintigraphy	Liver cancer	Liver biopsy with pathology	Index test positives (39%) and test negatives (63% were not verified for unspecified reasons.	(6)
D-dimer and alveolar dead space measurement	Pulmonary embolism	Pulmonary angiography	Not all patients were verified by angiography, for unspecified reasons.	(7)
Differential verification				
Index test(s)	Target condition	Reference standard	Problem	References
Elbow extension test	Elbow fracture	Radiography or follow up	Index test positives received radiography. Index test negatives received follow up.	(8)
D-dimer test	Deep-vein thrombosis (DVT)	Ultrasonography of the legs	Patients with negative d-dimer test or clinically low risk of DVT were verified by follow-up at 3 months.	(9)
Patient history, physical examination and laboratory tests	Serious bacterial infection	Cultures of blood, spinal fluid, urine, stool positive for a pathogen or a panel diagnosis	Mixture of reference standards, as used in clinical practice.	(10)
Ventilation/perfusion lung scans	Acute pulmonary embolism	Scintigraphy or pulmonary angiography or follow up	Mixture of reference standards, as used in clinical practice.	(11)

Partial verification

Often not all study subjects who underwent the index test receive the reference standard, leading to missing disease outcome data (Figure 1B). These situations have been labelled partial verification. The bias associated with partial verification is called partial verification bias, work-up bias, or referral bias.[12-14]

Different mechanisms can lead to partial verification, as will be illustrated by the following examples (see also Table 1). When target conditions produce lesions that need biopsy and subsequent histological verification as in many cancers, it is impossible to verify negative index test results ('where to biopsy?'). An example is FDG-PET scanning to detect possible distant metastases before planning major curative surgery in patients with carcinoma of the oesophagus: only PET hot spots can be biopted and verified histologically.[3]

Ethical reasons can also play a role in withholding a reference standard. Angiography is still considered the best available method for the detection of pulmonary embolisms. Because of its invasiveness and risk of serious complications it is now considered unethical to perform this reference standard in low-risk patients, such as those with a low clinical probability and negative d-dimer result.[7]

Sometimes the reference standard may be temporarily unavailable or patients and doctors can decide to refrain from disease verification. In a study evaluating the accuracy of digital rectal examination and prostate specific antigen (PSA) for the early detection of prostate cancer, 145 out of 1000 men fulfilled the criterion for verification by the reference standard: transrectal ultrasound combined with biopsy. Fifty-four of these men did not undergo the reference standard, for unknown reasons.[4] In another study, the accuracy of dobutamine-atropine stress echocardiography for the diagnosis of coronary artery disease was assessed using coronary angiography as reference.[5] Only a small part of patients received this reference standard because of the practitioners' decision to refer to angiography depended on the patient's history and test results.

The above examples show that partial disease verification and thus missing disease outcome status in part of the patients, is often not completely at random or a-selective. It is usually based on results of the index test under study or other observed patient variables or test results. If so, the missing outcome status is selectively missing, as the reason for disease verification is associated with other information. For example, patients with a positive index test result or patients with a high clinical suspicion based on various variables (i.e. high pre-index test probability) are often more likely to be verified by the reference test than patients with negative test results or a low pre-index test probability. Simply leaving

such selectively unverified patients out of the analysis will leave a non-random (selective) part of the subjects for the analysis, and thus generate biased estimates of the accuracy of the index test under study. The direction and magnitude of this bias will depend on how selective the reason for non-verification is, the number of patients that are unverified, and the ratio between the number of patients with positive and negative index test results that remain unverified.[14] The bias always occurs in the estimates of the sensitivity and specificity of the diagnostic index test or model under study, and often as well in the predictive value. When the reason for partially missing outcomes is only based on the results of the index test under study, the predictive values of this index test will indeed be unbiased (see below).

If, however, the reason for referral for reference testing is not only due to the index test results but also to other patient information, the predictive values of the index test will also be affected.[15]

One of the early methods to correct for partial verification bias was developed by Begg and Greenes.[16] In short, this method uses only the pattern of reference-standard-verified diseased and non-diseased, among the patients with a positive or negative result of the (single) index test under study. This pattern is then used to calculate the expected number of diseased and non-diseased among the non-verified patients with a positive or negative index test result, to obtain an inflated two-by-two table as if all patients were verified by the reference standard. This correction method assumes that the reason for referral to the reference test is only due to the result of the index test under study. Hence, conditional on these index test results, the decision to verify is in fact a random process. The method can also be extended to more than one test result, but this requires exact knowledge of the reasons and patterns behind the partial disease verification.[16;17]

More recently, multiple imputation methods have been proposed to correct for partial verification problems.[18;19] Multiple imputation can be viewed as a 'statistical' work-out of the intuitive 'diagnostic reasoning' of the clinician. Just as a clinician in practice refers a patient for disease verification by a (more invasive, burdening or costly) reference standard using all available patient information, multiple imputation techniques use also all available information of the patient - and that of similar patients - to estimate the most likely value of the missing reference test result in non-verified patients.

Imputation methods comprise two phases: an imputation phase where each missing reference test result is estimated and imputed, using all available patient information and an analysis phase where accuracy estimates of the diagnostic index test or model are computed by standard procedures, based on the now completed dataset. Several imputation

variants are available, ranging from single imputation of missing reference test values to multiple imputation[20;21]. Instead of filling in a single value for each missing value, as with single imputation, multiple imputation procedures replace each missing value with a set of plausible values to represent the uncertainty about the imputed value. These multiple imputed data sets are then analyzed, one by one, again by standard procedures. In a next step the results from these analyses are combined to produce accuracy estimates of the diagnostic index test(s) or model and their confidence intervals that properly reflect the uncertainty due to missing values.[20;21]

To optimally apply multiple imputation techniques to address partial verification, it is important for researchers to collect as much as possible detailed data on study subjects that could potentially drive the (selective) referral for reference testing. The performance of the multiple imputation or other correction methods will improve with more and better information that may be involved in disease verification decisions. The flexibility of the multiple imputation method enables the incorporation of multiple pieces of observed patient information, and not only the results of the index test under study, thereby increasing the likelihood of correctly imputing missing reference test values in patients in whom the disease status was selectively not verified by the reference standard.[17-19]

Finally, we stress that all discussed mathematical methods to correct for selectively missing disease-outcome status or reference test results, and thus for partial verification bias, make use of observed (patient) information or variables. They assume that the reasons for missingness depend on observed information only. Clearly, this assumption can not be tested with the data at hand, simply because non-observed information is by definition not available. If one expects selectively missing reference test results due to unobserved information, there are methods to perform additional (sensitivity) analysis to quantify to what extent the diagnostic accuracy estimates of the index test change under these situations.[22;23]

Differential verification

Another frequently encountered approach in diagnostic accuracy studies is to use an alternative, second best, reference test in those subjects where the result of the first, preferred reference test can or will not be obtained (Figure 1C). Although this seems a clinically appealing and ethical approach, bias arises when the results of the two reference tests are treated as interchangeable. Both reference tests are almost by definition of different quality in terms of target disease classification or may even define the target disease differently.[24;25] Hence, simply combining all disease-outcome data in a single analysis (Figure 2), as if both reference tests are yielding the same disease outcomes, does not validly reflect the 'true'

disease presence/absence status. The so-estimated disease prevalence differs from what one would have obtained if all subjects had undergone the preferred reference standard. Consequently, all estimated measures of accuracy of the diagnostic index test or model will be biased. This is called differential verification bias.[12;13]

(i) Verification with preferred reference test				(ii) Verification with alternative reference test				(iii) Differential verification with either reference test		
	R+	R-			S+	S-			?+	?-
T+	a	b	+	T+	-	-	≠	T+	a	b
T-	-	-		T-	c	d		T-	c	d

Figure 2. Diagnostic accuracy study with differential disease verification in which the preferred reference test R is used in only the index test positives (i), an alternative but less perfect reference test S used in the index test negatives (ii), which are then simply combined to form one overall complete two-by-two table ignoring the fact that both reference tests have different abilities to determine the disease presence/absence, such that the disease status in fact is ambiguously defined (iii).

For example, when evaluating a new marker for acute appendicitis, histopathology of the appendix is the preferred reference test, but clinical follow-up is sometimes used as an alternative reference test, e.g. if histopathology is considered too invasive for a patient. Compared to histopathology, clinical follow-up is likely to have a higher implicit threshold to detect appendicitis, so it will label more patients as *non-diseased* (i.e. no appendicitis). It illustrates that these two reference tests define the target condition in a different way. Histopathology seems the preferred reference test because it reveals even the smallest amount of inflamed cells. One could argue that the more relevant information for clinical practice is not whether the patient has any inflamed cells, but whether the patient will recover without intervention. This would make natural history the clinically preferred reference, even though it would be unethical to use follow up in all subjects and to withhold surgery. It does mean that accuracy estimates from a combination of histopathology and follow-up will systematically differ from what one would have obtained if *all* index test results had been verified by either clinical follow-up or histology. Because accuracy estimates of the new marker ignore the use of different reference tests, they are also difficult to interpret. In situations of differential verifications like this, the results should be corrected and reported separately for each reference standard, to provide informative and unbiased measures of accuracy of the diagnostic index test or model. We illustrate this using a clinical example from the recent literature[8] in Box 1.

Recently, a Bayesian method was proposed for simultaneously adjusting for differential verification bias and for the fact that these multiple reference tests were imperfect.[26] The method produces accuracy measures both with respect to the latent disease status and

with respect to the use of different reference tests. The former can be considered as a more general measure of performance of the index test with respect to a theoretically defined target condition or disease status as none of the used reference tests is considered as 'perfect'. However, the index tests' accuracy measures for each of the reference standards may be considered of greater and direct clinical relevance, as these reflect the accuracy against the reference tests that are commonly also performed in daily practice, on which further patient management decisions will often be based.

Conclusion

In diagnostic accuracy studies, all efforts should be made to verify as many patients as possible – preferably all - with the optimal reference test, to avoid bias. In practice, patient burden, costs or other reasons may often prevent this from happening (Table 1).[27]
If the disease-outcome status is verified by the reference test in only part of the patients, which is usually selective disease verification based on other observed patient information, we advise to use the above described mathematical correction methods to correct for the partial verification bias.[16;17;19]

There is yet insufficient knowledge and evidence to make general statements about what proportion of missing reference standard results might be acceptable and at which point correction methods will become unreliable. Following various statistical guidelines,[18-21;28;29] we recommend using these correction methods even with small rates of missing disease-outcomes. Even small proportions of missing outcomes may yield biased accuracy estimates of the index test(s) or model under study, if the non-verified sample is highly selective.

Which upper limits of missing reference test data can still be corrected for, is even harder to say.[13] Recently, Janssen et al showed that even for large amounts of missing data, imputation leads to less biased results than simply ignoring the (selectively) non-measured subjects.[28] The authors discussed that this possibility for imputation depends on how selective or different the observed versus non-observed subjects are, and how much subjects or data is left to build well-enough imputation models. In any case, authors applying correction or imputation methods for addressing partial verification methods should provide insight in both issues: how many subjects had missing reference test values and how different were the verified versus non-verified patients by comparing both groups on the observed characteristics.[29;30]

If the preferred reference is not possible and thus missing in complete subgroups (e.g. in index test negatives or with a low pre-test probability), applying a different usually less perfect 'reference test' will obviously produce different information about the disease status.

Box 1. Clinical Example: the Elbow Extension Test to rule out Elbow Fracture

In a recent study[16] the Elbow Extension Test (EET) was studied on its accuracy to rule out elbow fractures. The preferred reference test was radiography. For unstated reasons (costs, efficiency or radiation reduction), radiography was planned in patients with a positive EET result and the negative EET patients received a structured follow-up assessment by telephone after 7-10 days to verify whether elbow fracture was indeed absent (the alternative reference test). Only patients who met any of the pre-specified recall criteria were asked to return to the emergency department for radiography after all. The rest were considered not to have a (clinically significant) elbow fracture. The resulting data are shown in Table 2.

Table 2

	Radiography		Follow Up	
	fracture	no fracture	fracture	no fracture
Elbow extension test +	521	617	NA	NA
Elbow extension test -	14 †	167 †	3	414

† Data available due to 'protocol violations'

The authors reported overall estimates of accuracy of the EET, ignoring the use of different reference standards (Table 3, first row). Though both radiography and structured follow-up are useful verification methods, their results are not necessarily interchangeable.

Table 3 Corrected sensitivities, specificities and predictive values (in percentages) of the Elbow Extension Test

Analysis	Sensitivity (CI)	Specificity (CI)	NPV(CI)	PPV (CI)
Differential verification ignored*	96.8 (95.0 to 98.2)	48.5 (45.6 to 51.4)	97.2 (95.5 to 98.3)	45.8 (42.9 to 48.7)
Corrected for partial verification (accuracy wrt to radiography)				
Begg & Greenes	91.8 (88.0 to 95.7)	47.2 (44.2 to 50.2)	92.3 (88.6 to 95.9)	45.8 (42.6 to 49.0)

NPV = Negative Predictive Value; PPV = Positive Predictive Value; CI = 95% Confidence Interval; wrt = with respect to

The availability of 181 negative EET patients who were after all evaluated by the preferred reference test ('protocol violations': Table 2), enables to apply the above mentioned correction methods for partial verification, under the assumption that conditional on the index test result, the decision to verify can be seen as a random process.

The corrected values of sensitivity and specificity clearly illustrate the consequences of differential verification (Table 3, second row). We found differences in the estimates of EET accuracy when verification bias is simply ignored and when it is adjusted for.

The negative predictive value (which was of primary interest: to rule out elbow fractures, aiming at an NPV of at least 97%), with respect to radiography alone, was lower than the negative predictive value reported by the authors. This clearly shows that two reference tests should not be viewed as one without concern.

For a more detailed and elaborate discussion of this example and the possibilities to correct for differential verification, we refer to the recently published: 'Adjusting for differential verification bias in diagnostic accuracy studies: A Bayesian approach'.[26]

Because overall accuracy estimates of the diagnostic index test or model under study that ignore the use of different reference tests are difficult to interpret, the results should be reported separately for each 'reference test' to provide more clinically informative and indeed unbiased measures of diagnostic accuracy.[12] If in these situations one still wants to quantify the accuracy of the diagnostic index test or model with regard to the same underlying target condition, one should also correct for possible imperfections of the applied reference tests.[26]

Reference List

1. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, editor. The Evidence Base of Clinical Diagnosis. 2nd ed. London: BMJ Books; 2002. 39-60.
2. Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic procedures. The evidence base of clinical diagnosis. London: BMJ Books; 2002. 1-18.
3. Lee J, Aronchick JM, Alavi A. Accuracy of F-18 fluorodeoxyglucose positron emission tomography for the evaluation of malignancy in patients presenting with new lung abnormalities: a retrospective review. Chest 2001; 120(6):1791-1797.
4. Pode D, Shapiro A, Lebensart P, Meretyk S, Katz G, Barak V. Screening for prostate cancer. Isr J Med Sci 1995; 31(2-3):125-128.
5. Elhendy A, van Domburg RT, Poldermans D, Bax JJ, Nierop PR, Geleijnse ML et al. Safety and feasibility of dobutamine-atropine stress echocardiography for the diagnosis of coronary artery disease in diabetic patients unable to perform an exercise stress test. Diabetes Care 1998; 21(11):1797-1802.
6. Drum DE, Christacopoulos JS. Hepatic scintigraphy in clinical decision making. J Nucl Med 1972; 13(12):908-915.
7. Kline JA, Israel EG, Michelson EA, O'Neil BJ, Plewa MC, Portelli DC. Diagnostic accuracy of a bedside D-dimer assay and alveolar dead-space measurement for rapid exclusion of pulmonary embolism: a multicenter study. JAMA 2001; 285(6):761-768.
8. Appelboam A, Reuben AD, Benger JR, Beech F, Dutson J, Haig S 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.
9. Buller HR, Ten Cate-Hoek AJ, Hoes AW, Joore MA, Moons KG, Oudega R et al. Safely ruling out deep venous thrombosis in primary care. Ann Intern Med 2009; 150(4):229-235.
10. Bleeker SE, Moons KG, rksen-Lubsen G, Grobbee DE, Moll HA. Predicting serious bacterial infection in young children with fever without apparent source. Acta Paediatr 2001; 90(11):1226-1232.

11. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). The PIOPED Investigators. JAMA 1990; 263(20):2753-2759.
12. Rutjes AW, Reitsma JB, Irwig LM, Bossuyt PM. Sources of bias and variation in diagnostic accuracy studies. In: AWS Rutjes, editor. Partial and differential verification in diagnostic accuracy studies. Amsterdam: Rutjes; 2005. 31-44.
13. 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):iii, ix-51.
14. 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(8):797-806.
15. Little RA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 1987.
16. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39(1):207-215.
17. de Groot JA, Janssen KJ, Zwinderman AH, Bossuyt PM, Reitsma JB, Moons KG. Correcting for partial verification bias: a comparison of methods. Ann Epidemiol 2011; 21(2):139-148.
18. Harel O, Zhou XH. Multiple imputation for correcting verification bias. Stat Med 2006; 25(22):3769-3786.
19. de Groot JA, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB. Multiple imputation to correct for partial verification bias revisited. Stat Med 2008; 27(28):5880-5889.
20. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59(10):1087-1091.
21. Rubin DB. Multiple imputation for non response in surveys. New York: Wiley; 1987.
22. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. Biometrics 2003; 59(1):163-171.
23. Kosinski AS, Barnhart HX. A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. Stat Med 2003; 22(17):2711-2721.
24. 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(3):189-202.
25. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11):1061-1066.

26. de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Bossuyt PM, Moons KG. Adjusting for Differential-verification Bias in Diagnostic-accuracy Studies: A Bayesian Approach. Epidemiology 2011; 22(2):234-241.
27. Oostenbrink R, Moons KG, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. J Clin Epidemiol 2003; 56(6):501-506.
28. Janssen KJ, Donders AR, Harrell FE, Jr., Vergouwe Y, Chen Q, Grobbee DE et al. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 2010; 63(7):721-727.
29. Mackinnon A. The use and reporting of multiple imputation in medical research - a review. J Intern Med 2010; 268(6):586-593.
30. van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006; 59(10):1102-1109.

Chapter 3

Partial verification bias

Chapter 3.1

Multiple imputation to correct for partial verification bias: A revision of the literature

Joris A.H. de Groot
Kristel J.M. Janssen
Koos A.H. Zwinderman
Karel G.M. Moons
Johannes B. Reitsma

Stat Med 2008; 27(28):5880-5889

Abstract

Partial verification refers to the situation where a subset of patients is not verified by the reference (gold) standard and is excluded from the analysis. If partial verification is present, the observed (naive) measures of accuracy such as sensitivity and specificity are most likely to be biased. Recently, Harel and Zhou showed that partial verification can be considered as a missing data problem and that multiple imputation (MI) methods can be used to correct for this bias. They claim that even in simple situations where the verification is random within strata of the index test results, the so-called Begg and Greenes (B&G) correction method underestimates sensitivity and overestimates specificity as compared with the MI method. However, we were able to demonstrate that the B&G method produces similar results as MI, and that the claimed difference has been caused by a computational error. Additional research is needed to better understand which correction methods should be preferred in more complex scenarios of missing reference test outcome in diagnostic research.

Introduction

Partial verification is one of the major problems in the assessment of the accuracy measures of diagnostic tests. This problem occurs when only a sub sample of those patients who are initially tested subsequently receives the definitive assessment for disease status (e.g. verification by the reference standard). If partial verification is present, the observed (Naive) measures of accuracy like sensitivity (the probability of a positive test given the true disease status is positive) and specificity (the probability of a negative test given the true disease status is negative) are biased. Several solutions have been proposed to correct or to alleviate this bias.[1]

One of the early and frequently applied methods to correct for this type of bias was developed by Begg and Greenes (B&G),[2] based on the assumption of conditional independence. This means that conditional on the result of the test under evaluation, the mechanism for selecting the sample for verification can be considered to be random (ignorable verification assumption).

Recently Harel and Zhou[3] described that partial verification can be considered as a missing data problem and that Multiple Imputation (MI) methods can be used to correct for this bias. They claim that the so-called Begg and Greenes (B&G) correction method is underestimating sensitivity and overestimating specificity as compared to the MI method, even in the simple case where the pattern of missing values is only determined by a single factor.

They use the Diaphanography data for breast cancer (Table 1) to illustrate their point that MI produces different results than the B&G method. Therefore, they recommend the use of MI when analysing a dataset from a diagnostic accuracy study where partial verification is present.

Table 1. Data structure of Diaphananography for breast cancer example as used by Harel and Zhou[3]

Cross table			
		T=1	T=0
V=1	D=1	26	7
	D=0	11	44
V=0		30	782
Total		67	833

D=1 if disease present; D=0 if disease not present; T=1 if positive test result; T=0 if negative test result; V=1 if verified by golden standard procedure; V=0 if not verified by golden standard procedure

Hanley and Begg[4] already commented that the B&G method can be considered as a straightforward 'single imputation' based on exactly the same (ignorability) assumption that Harel and Zhou have used for their MI techniques. Therefore, it would be unrealistic that the B&G and MI methods could produce such starkly different estimates as those reported by Harel and Zhou,[3] unless a computational error had been made. Harel and Zhou in return answered that this however was not the case.[5]

Methods and Results

Prompted by this mutual correspondence, we have repeated the analysis as described by Harel and Zhou in their original paper.[3] We also found striking results, indicating that indeed the claims made by Harel and Zhou may be less tenable. Our conclusion is based on the following observations.

First we used the corrected estimates of sensitivity and specificity by Harel and Zhou to recalculate the prevalence of the disease to check the plausibility of their estimates.

With D being disease status and T being the index test result one can write the probability of observing a positive test result as:

$$P(T^+) = P(T^+ \mid D^+) \times P(D^+) + P(T^+ \mid D^-) \times P(D^-) \qquad (1)$$

By rearranging equation (1) we can write disease prevalence as a function of $P(T^+)$, sensitivity (sens) and specificity (spec)

$$P(D^+) = (P(T^+) + spec - 1) / (sens + spec - 1) \qquad (2)$$

Because in the Diaphonagraphy cohort study every patient received the test under evaluation, the estimate of $P(T^+)$ is unbiased irrespective of partial verification and we can apply the estimated sensitivity and specificity of both the B&G method and the Rubin MI method in formula (2) to obtain an estimate of the prevalence. For the B&G method we obtain:

$$P(D^+) = (0.074 + 0.974 - 1) / (0.301 + 0.974 - 1) = 0.175$$

This estimate seems to be a valid and plausible value given the disease under study. But using the estimators of the Rubin MI method (and thus all other reported MI estimators) we obtain:

$$P(D^+) = (0.074 + 0.862 - 1) / (0.714 + 0.862 - 1) = -0.111$$

Obviously, negative disease prevalence is impossible indicating that the reported MI estimators for sensitivity and specificity appear to be invalid.

Secondly, we repeated the MI analysis with 10 imputed datasets (Appendix 1) using Mice,[6] in an attempt to reproduce the results by Harel and Zhou (Table 2).

Table 2. Observed (Naive) and corrected estimates of sensitivity and specificity according to Harel and Zhou[3]

Correction method	Sensitivity	Specificity
Naive	0.788	0.800
B&G	0.292	0.973
Rubin MI	0.714	0.862

Naive: complete case analysis; B&G: Begg and Greenes correction method; Rubin MI: Rubin's Multiple Imputation correction method

In contrast to Harel and Zhou, our estimates of sensitivity and specificity of the MI method did not differ much from the results of the B&G method. We then repeated the analysis using aregImpute[7] (appendix 1) and the MI procedure in SAS[8] (Appendix 2) and found similar results (Table 3).

Table 3. Replication of the observed (Naive) and corrected estimates of sensitivity and specificity

Correction Method	Sensitivity	Specificity
Naive	0.788	0.800
B&G	0.292	0.973
MI mice	0.301	0.974
MI aregImpute	0.296	0.973
MI SAS	0.298	0.974

Naive: complete case analysis; B&G: Begg and Greenes correction method; MI mice: Multiple Imputation correction method using mice; MI aregImpute: Multiple Imputation correction method using aregImpute; MI SAS: Multiple imputation correction method using SAS

Prompted by these results, we explored the 10 individual imputed data sets to study the variability in results of sensitivity and specificity (Appendix 3) and did a remarkable finding. When it became clear that the range in sensitivity and specificity could not account for the found differences in results we calculated the mean positive predictive and negative predictive values from the 10 imputed data sets. Our calculated value for the positive predictive value (0.715) was strikingly similar to the reported value for sensitivity (0.714) by Harel and Zhou, whereas our calculated value for the negative predictive value (0.866) showed great similarity with their reported value of specificity (0.862).
Therefore, it is likely that the large difference in results between the MI and B&G reported by Harel and Zhou is due to a computational error.

Concluding remarks

In contrast to the conclusion of Harel and Zhou,[3] our study shows that the Begg and Greenes method leads to similar results as Multiple Imputation.

We do stress that both methods assume that the mechanism leading to missing values are known and observed, whereas in practice these mechanisms are not always known completely. We agree that Multiple Imputation may be a more flexible approach for incorporating various covariates that are related to missing values. Additional research is needed to better understand which correction methods should be preferred in various scenarios of missing data.

References

1. Reitsma JB, Rutjes AWS, Khan KS, Coomarasamy A, Bossuyt P. NHS Health Technology Assessment Programme: A systematic review of methods to evaluate tests when there is no gold standard. URL: http://www.hta.ac.uk/project/1573.asp Accessed on November 28, 2007.
2. Begg CB, Greenes RA. Assessment of Diagnostic Tests When Disease Verification is Subject to Selection Bias. *Biometrics* 1983; 39 : 207-215.
3. Harel O, Zhou XH. Multiple imputation for correcting verification bias. *Statistics in Medicine* 2006; 25 : 3769-86.
4. Hanley JA, Dendukuri N, Begg CB. Multiple imputation for correcting verification bias by Ofer Harel and Xiao-Hua Zhou, Statistics in Medicine 2006; 25:3769-3786. *Statistics in Medicine* 2006; 26 : 3046-47.
5. Harel O, Zhou XH. Rejoinder to Multiple imputation for correcting verification bias. *Statistics in Medicine* 2006; 26 : 3047-50.
6. Buuren S van, Oudshoorn K. Flexible multivariate imputation by mice. Technical report. Leiden, The Netherlands: TNO prevention and Health. http://www.stefvanbuuren.nl/publications/Flexible%20multivariate%20-%20TNO99054%201999.pdf Accessed on November 28,2007.
7. Harrell FE. The Hmisc library, 2002. http://hesweb1.med.virginia.edu/biostat/s/Hmisc.html Accessed on October 31, 2007.
8. SAS software, Version 9.1 of the SAS System for Windows. Copyright © 2007 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA. http://www.sas.com/ Accessed on November 28, 2007.

Appendix 1. R-script Multiple Imputation, using mice and aregImpute

```
library(Hmisc)
library(Design)
library(foreign)
library(mice)

#Simulating the diaphananography data used by Harel and Zhou
T                   <- c(rep(1,26),rep(1,11),rep(0,7),rep(0,44),rep(1,30),rep(0, 782))
D                   <- c(rep(1,26),rep(0,11),rep(1,7),rep(0,44),rep(NA,30),rep(NA,782))
ddata               <- data.frame(cbind(T, D))

#Imputation using Mice
ddataImputm         <- mice(ddata,m=10)
tm                  <- complete(ddataImputm, action="repeated")
sensimpm <- specimpm <- ppvm <- npvm <- rep(NA,10)

#Loop to calculate the sensitivity, specificity, positive predictive values and negative #predictive values in the ten imputed datasets
for (i in 1:10)
{
imptm               <- table (tm[,10+i],tm[,1+i])
sensimpm[i]         <- imptm[2,2] / (imptm[2,2] + imptm[2,1])
specimpm[i]         <- imptm[1,1] / (imptm[1,1] + imptm[1,2])
ppvm[i]             <- imptm[2,2] / (imptm[2,2] + imptm[1,2])
npvm[i]             <- imptm[1,1] / (imptm[1,1] + imptm[2,1])
}

#Calculating the mean sensitivity, specificity, positive predictive value and negative #predictive value
sensitivitym        <- mean (sensimpm)
specificitym        <- mean(specimpm)
ppvm                <- mean(ppvm)
npvm                <- mean(npvm)


#Imputation using aregImpute
ddataImputa         <- aregImpute(~T+D, n.impute=10,data=ddata)

# The aregComplete function to extract the imputed datasets
aregComplete        <- function (xtrans, data, impnr)
{
using.Design        <- FALSE
used.mice <- any(oldClass(xtrans) == "mids")
if (used.mice)
completed.data      <- complete(xtrans, impnr)
else {
completed.data      <- data
imputed.data        <- impute.transcan(xtrans, imputation = impnr,
data = data, list.out = TRUE, pr = FALSE, check = FALSE)
completed.data[names(imputed.data)] <- imputed.data
}
completed.data
}
ta                  <- matrix(NA,900,10)
for (i in 1:10)
{
ta[,i]              <- aregComplete(ddataImputa ,ddata,i)[,2]
```

```
}
sensimpa  <- specimpa <- ppva <- npva <- rep(NA,10)

# Loop to calculate the sensitivity, specificity, positive predictive values and negative
# predictive values in the ten imputed datasets.
for (i in 1:10)
{
impta                <- table (ta[,i],ddata[,1])
sensimpa[i]          <- impta[2,2] / (impta[2,2] + impta[2,1])
specimpa[i]          <- impta[1,1] / (impta[1,1] + impta[1,2])
ppva[i]              <- impta[2,2] / (impta[2,2] + impta[1,2])
npva[i]              <- impta[1,1] / (impta[1,1] + impta[2,1])
}
sensitivitya         <- mean(sensimpa)
specificitya         <- mean(specimpa)
ppva                 <- mean(ppva)
npva                 <- mean(npva)

# Imputation results
results              <- matrix(NA, 2,4)
colnames(results)    <- c("sensitivity","specificity","ppv","npv")
rownames(results)    <- c("mice","aregImpute")
results[1,1] <- sensitivitym; results[2,1] <- sensitivitya; results[1,2] <- specificitym;
results[2,2] <- specificitya;        results[1,3] <- ppvm; results[2,3] <- ppva; results[1,4] <- npvm; results[2,4] <-
npva
round(results, 3)
```

Appendix 2. SAS-script Multiple Imputation

```
* Construct dataset: Diaphanography for breast cancer ;
data partial;
        do i=1 to 26;
                d=1; t=1; v=1;          output;
        end;
        do i=1 to 7;
                d=1; t=0; v=1;          output;
        end;
        do i=1 to 44;
                d=0; t=0; v=1;          output;
        end;
        do i=1 to 11;
                d=0; t=1; v=1;          output;
        end;
        do i=1 to 30;
                d=.; t=1; v=0;          output;
        end;
        do i=1 to 782;
                d=.; t=0; v=0;          output;
        end;
run;

* Impute D outcome using only T result ;
* Imputation method: monotone logistic, 10 rounds ;

proc mi data=partial seed=1230 nimpute=10 out=impute;
        class d;
        monotone logistic(d=t / details);
        var t d;
run;

* Calculate sensitivity in each imputed dataset (m=10) ;
ods output binomialprop=sens_impute;
proc freq data=impute(where=(D eq 1)) order=data;
        by _imputation_;
  tables t / binomial;
run;
ods output close;

* Calculate specificity in each imputed dataset (m=10) ;
ods output binomialprop=spec_impute;
proc freq data=impute(where=(D eq 0)) order=data;
        by _imputation_;
  tables t / binomial;
run;
ods output close;



* Select sensitivity and correspoding SE of each imputed datatset in single row ;
data sens_ana;
        set sens_impute;
        by _imputation_;
        retain sens se;
```

```
        if first._imputation_ then do;
                sens=.; se=.;
        end;
        if name1 eq '_BIN_' then sens=nvalue1;
        else if name1 eq 'E_BIN' then se=nvalue1;
        if last._imputation_ then output;
        keep _imputation_ sens se;
run;

* Combine estimates from imputed datasets ;
proc mianalyze data=sens_ana;
  modeleffects sens;
  stderr se;
run;

* Select specificity and correspoding SE of each imputed datatset in single row ;
data spec_ana;
        set spec_impute;
        by _imputation_;
        retain spec se;
        if first._imputation_ then do;
                spec=.; se=.;
        end;
        if name1 eq '_BIN_' then spec=nvalue1;
        else if name1 eq 'E_BIN' then se=nvalue1;
        if last._imputation_ then output;
        keep _imputation_ spec se;
run;

* Combine estimates from imputed datasets ;
proc mianalyze data=spec_ana;
  modeleffects spec;
  stderr se;
run;
```

Appendix 3. Accuracy measures per imputed dataset, using mice

Imputed datasets				Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
		T=1	T=0				
Dataset 1	D=1	48	103	0.318	0.975	0.716	0.876
	D=0	19	730				
		T=1	T=0				
Dataset 2	D=1	45	96	0.319	0.971	0.672	0.885
	D=0	22	737				
		T=1	T=0				
Dataset 3	D=1	52	126	0.292	0.979	0.776	0.849
	D=0	15	707				
		T=1	T=0				
Dataset 4	D=1	47	104	0.311	0.973	0.701	0.875
	D=0	20	729				
		T=1	T=0				
Dataset 5	D=1	52	125	0.294	0.979	0.776	0.850
	D=0	15	708				
		T=1	T=0				
Dataset 6	D=1	46	111	0.293	0.972	0.687	0.867
	D=0	21	722				
		T=1	T=0				
Dataset 7	D=1	47	132	0.263	0.972	0.701	0.842
	D=0	20	701				
		T=1	T=0				
Dataset 8	D=1	45	101	0.308	0.971	0.672	0.879
	D=0	22	732				
		T=1	T=0				
Dataset 9	D=1	48	103	0.318	0.975	0.716	0.876
	D=0	19	730				
		T=1	T=0				
Dataset 10	D=1	49	811	0.293	0.975	0.731	0.858
	D=0	18	715				
Mean				0.301	0.974	0.715	0.866

D=1 if disease present; D=0 if disease not present; T=1 if positive test result; T=0 if negative test result

Chapter 3.2

Adjusting for partial verification bias in diagnostic accuracy studies: A comparison of methods

Joris A.H. de Groot
Kristel J.M. Janssen
Koos A.H. Zwinderman
Patrick M.M. Bossuyt
Johannes B. Reitsma
Karel G.M. Moons

Ann Epidemiol 2011; 21(2):139-148

Abstract

Background

A common problem in diagnostic research is that the reference standard has not been performed in all patients. This partial verification may lead to biased accuracy measures of the test under study. The authors studied the performance of multiple imputation and the conventional correction method proposed by Begg and Greenes under a range of different situations of partial verification.

Methods

In a series of simulations, using a previously published Deep Venous Thrombosis dataset (N=1292), the authors set the outcome of the reference standard to missing based on various underlying mechanisms and by varying the total number of missing values. They then compared the performance of the different correction methods.

Results and Conclusions

The results of the study show that when the mechanism of missing reference data is known, accuracy measures can easily be correctly adjusted using either the Begg and Greenes method, or multiple imputation. In situations where the mechanism of missing reference data is complex or unknown, we strongly recommend using multiple imputation methods to correct. These methods can easily apply for both continuous and categorical variables, are readily available in statistical software and give reliable estimates of the missing reference data.

Introduction

In studies of diagnostic accuracy, results from one or more tests under evaluation are compared with the results obtained with the reference standard. These studies are a vital step in the evaluation of new and existing diagnostic technologies. The reference standard is the best available method for identifying patients as having the disease of interest. Measures, such as sensitivity, specificity and predictive values, express how well tests under evaluation are able to identify patients as having the target disease.[1]

A common problem in diagnostic research is that the reference standard has not been carried out in all patients because of ethical, practical or other reasons. Partial verification, if not accounted for, is known to lead to biased accuracy estimates, described in the literature as partial verification bias or work-up bias.[2]

In clinical practice different mechanisms can lead to partial verification.[3] Sometimes it is simply unavoidable. For example, to verify results of Positron Emission Tomography (PET) in staging oesophageal cancer,[4] only results of patients with PET lesions suggestive of distant metastases can be verified by histology. Histology cannot be carried out in PET negative patients. Second, incomplete verification can be prespecified in the design, for example, for efficiency reasons. This is often the case in screening test evaluation studies, where disease prevalence is low.[5] In these types of studies, researchers often decide to apply the reference standard in only a random sample of the large group of patients with a negative screening test result. In other studies, partial verification is not planned, and reasons are unclear and not documented. For example, the accuracy of dobutamine atropine stress echocardiography for detecting coronary artery disease can be assessed using coronary angiography as the reference standard. In one study,[6] only a small sample of the patients received this reference standard because of the practitioners' decision to refer patients to angiography or not, depending on history and other test results.

One of the methods to correct for partial verification was developed by Begg and Greenes (B&G).[7] In short, this method uses observed proportions of diseased and non-diseased among the verified patients to calculate the expected number of diseased and non-diseased among non-verified patients. The two are combined to obtain a complete two-by-two table, as if all patients had received the reference standard. (for details see Appendix 1) This correction method requires knowledge about the reasons responsible for partial verification. It is disputable whether this correction method also leads to valid results when the reasons for partial verification are less clear-cut.

Recently Harel and Zhou[8] have shown that partial verification can be considered as a missing data problem and that Multiple Imputation (MI) methods, the practice of 'filling in' missing

data with plausible values, can be used to correct for this bias. Their conclusion that multiple imputation is generally better than the existing methods with regard to alleviating the bias and correcting confidence interval width has been debated.[9,10] Hanley et al[9] stated that the numerical differences between the B&G method and MI found by Harel and Zhou[8] were highly unlikely. De Groot et al[10] concluded that these differences were due to a computational error and therefore led to spurious conclusions.

We will compare the performance of multiple imputation and the correction method of B&G under a range of situations of partial verification using a simulation study and examine under which circumstances they produce similar results and when their results differ. Based on our findings we will propose guidance for researchers designing and analyzing diagnostic accuracy studies with partial verification.

Methods

We have used a previously published dataset, in which all patients had been verified by the reference standard. In a series of simulations, we deliberately set the outcome of the reference standard to missing based on various underlying mechanisms and by varying the total number of missing values, generating different partial verification patterns. We then compared the performance of different correction methods in each of these patterns of verification, in particular their ability to reduce the bias in estimates of accuracy by comparing it with the true value in the complete dataset.

Empirical dataset with complete verification

Data of a large study among adults with suspected Deep Venous Thrombosis (DVT) were used. For specific details of the study we refer to the literature.[11] In brief, 1292 consecutive patients with suspected DVT were included. DVT suspicion was primarily based on the presence of swelling, redness, or pain in one of the legs. After informed consent, the physician systematically documented the patient's history and the results of a physical examination. Subsequently, venous blood was drawn to measure D-dimer level. All patients were then referred to a hospital to undergo repeated compression ultrasonography of the lower extremities, which was used as the reference standard to determine the presence or absence of DVT. Repeated compression ultrasonography revealed DVT in 251 (19%) patients, of which 225 (90%) had a positive D-dimer test result.

In our series of simulations we used the complete data of 1292 research subjects (Table 1), to which we will refer as the original study group.

Table 1. Univariate Association of each Significant Diagnostic Variable with the Presence or Absence of DVT. Values are Percentages

Diagnostic variables	DVT present n=251 %	DVT absent n=1041 %
Patient history		
Gender + OC use:		
Males	47.4	36.2
Females using OC	12.0	9.2
Females not using OC	40.6	54.6
Absence of leg trauma	88.0	83.3
Presence of malignancy	6.8	3.3
Recent surgery	12.7	10.1
Physical examination		
Vein distention	19.5	15.5
Calf difference ≥ 3cm	62.9	34.3
Additional testing		
D-dimer abnormal ≥ 1000 ng/ml	89.6	39.5

DVT = deep venous thrombosis; n = number of patients; OC = oral contraceptive

D-dimer test results were dichotomized, labeling results as positive if they exceeded the 1000 ng/ml threshold. The reference test used in this study (repeated compression ultrasonography) was assumed to be 100% sensitive and 100% specific. The sensitivity, specificity and predictive values of the D-dimer test in the original study group were then calculated using standard methods.[1;12] These accuracy measures in the original study group will be referred to as the "true" sensitivity, specificity and predictive values. Confidence intervals were calculated using the Wilson "score" method.[13,14]

Patterns of partial verification

We selected a range of situations in which partial verification could typically arise in practise (Figure 1).

In the first pattern of missing values, the outcome of the reference standard was set to missing in a random subset of patients with a negative D-dimer test result. This reflects a common practical situation where the practitioner thinks it unnecessary to refer all subjects with a negative D-dimer test result to the hospital to undergo repeated compression ultrasonography. (Figure 1A) In this situation, the reason for practitioners to refer patients to the reference standard depends on a single variable known by the researcher: the D-dimer test result.

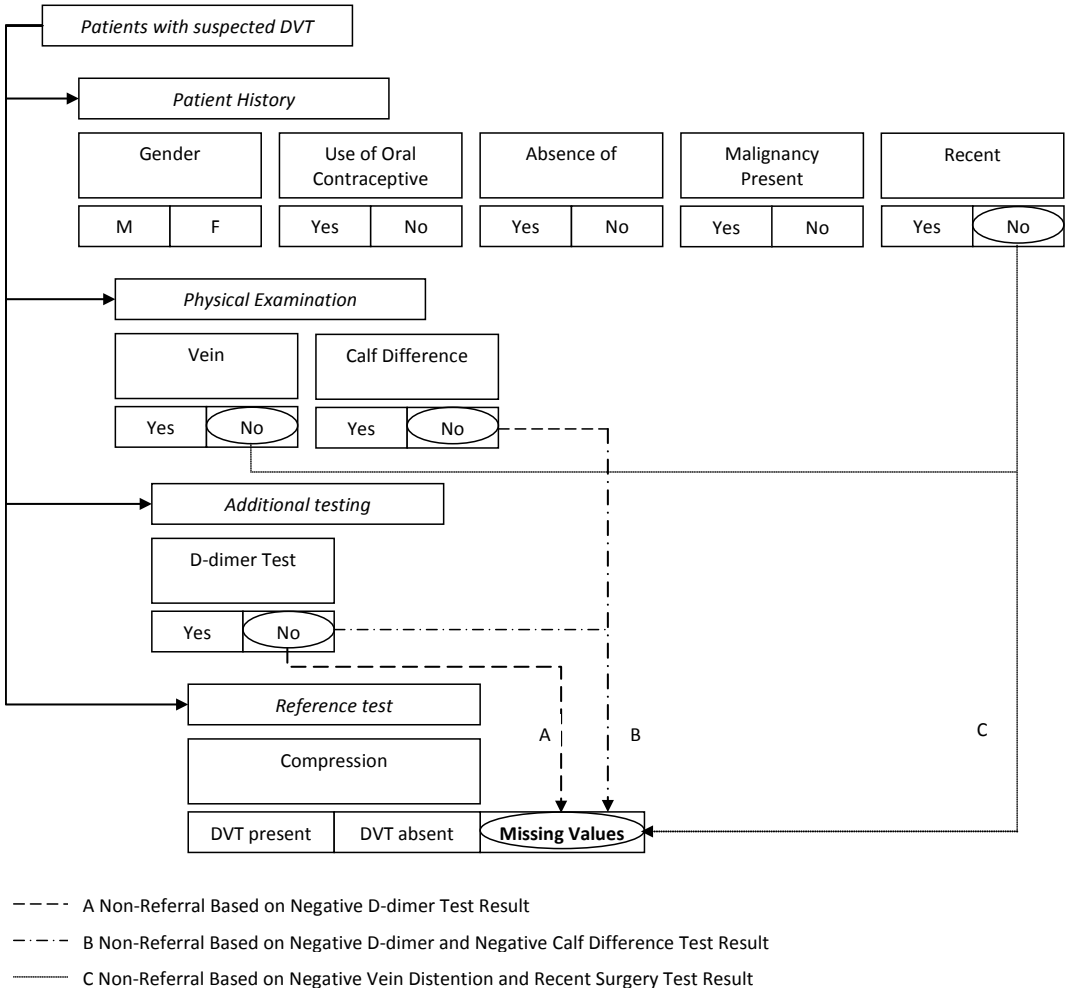

Figure 1. Three Different Mechanisms for Creating Missing Reference Standard Outcomes.

In the second partial verification pattern, the reference standard was set to missing depending not only on the result of the test under study but also on one other test result: the difference in calf circumference between the two legs. In this scenario, the practitioner decides that verification with compression ultrasonography is not necessary if both the D-dimer test result and the result of the Calf Difference test are negative (Figure 1B).

In the third pattern, the reference standard was set to missing depending on the results of two other test results: the absence of vein distension in the legs and the absence of recent surgery. This situation reflects the case that the practitioner combines several pieces of information to evaluate the likelihood of DVT. Practitioners will not carry out ultrasonography in all patients who did not undergo surgery recently and did not show signs of vein distension, regardless the D-dimer test result (Figure 1C).

For each of these three patterns, we introduced missing values for the outcome of the reference standard in 10% (n=129), 20% (n=258) and 30% (n=387) of all 1292 subjects, generating 9 scenarios with missing reference standard information. Missing values for the outcome in 10%, 20% and 30% of all the subjects correspond to:

- 19.7%,39.4%, and 59.1% of the subset of patients with a negative D-dimer test result (pattern A);
- 28.2%, 56.4%, and 84.6% of the subset of patients with a negative result on both the D-dimer test and the calf circumference (pattern B); and
- 13.4%, 26.9%, and 40.3% of the subset of patients who did not undergo surgery recently and did not show signs of vein distension (pattern C).

For each scenario, we generated 100 datasets, using random sampling with replacement.

Correction methods

In each dataset, we calculated the naive, complete case accuracy measures for D-dimer, as well as corrected accuracy measures, using both the B&G method and MI.

1. Complete case analysis

   Using this method, all nonverified cases were omitted from the analysis. For the remaining complete cases the accuracy measures, standard errors and 95% confidence intervals of the test under study were calculated in the standard way.[1;12]

2. B&G correction method

   This method was developed to improve the accuracy of sensitivity, specificity and predictive values in the case of missing data compared to the complete case analysis. Begg and Greenes proposed formulas[7] to first inflate the numbers in the two-by-two table, under the assumption that within the strata of the dependent variables the distribution is random, and then compute the accuracy measures. (Appendix 1.) Their formulas to calculate appropriate standard errors were also used in this study.[7] In our analyses, we formulated the B&G method in two different ways. First, only the index test under study (D-dimer) was considered as the cause of the missing reference data (B&G1; Appendix 1.1). Second, both the index test under study as well as the difference in calf circumference was considered as the cause of the missing reference data (B&G2; Appendix 1.2).

3. Multiple Imputation

In the imputation process all variables significantly associated with the reference standard (i.e., all variables in Table 1.) were used to compose the logistic regression model. The D-dimer result was included into the regression model as a dichotomous variable. We did not add interaction terms in the model because on clinical grounds we could not expect any major effect modification given the nature of the individual components in the model, and conform various diagnostic rules in the field of DVT constructed in the past decades (without interaction terms). We used this imputation model to create 10 imputed datasets.[15] For these imputed datasets, D-dimer accuracy measures were calculated.[1;12] The results were than combined using standard statistical methods in a way that reflects the extra variability due to missing data.[15]

Data analysis

In each dataset we calculated estimates, based on complete cases, and corrected estimates using the two approaches mentioned. We calculated the means of these estimates across the 100 simulations. We compared estimated sensitivities, specificities and predictive values of the D-dimer test and their respective 95% confidence intervals for all correction methods using figures and accompanying mean-square errors. All analyses were done using R 2.5.0 (R Foundation for Statistical Computing, www.R-project.org). The simulation scripts are available on request.

Results

Standard calculations of sensitivity, specificity and predictive values resulted in "true" accuracy measures of the original study group (Table 2).

Table 2. True Accuracy Measures of the D-dimer Test in the Original Study Sample with Standard Errors and 95% Confidence Intervals.

D-dimer	Estimate	SE	95% CI
Sensitivity	0.896	0.019	0.859-0.933
Specificity	0.605	0.015	0.576-0.634
Positive Predictive Value	0.354	0.019	0.317-0.391
Negative Predictive Value	0.960	0.008	0.944-0.976

SE = Standard Error; 95% CI = 95% Confidence Interval

With nonreferral based on negative D-dimer test result (Figure 2) an increasing deviation from the true values for sensitivity and specificity is seen with complete case analysis. Sensitivity is increasingly overestimated whereas specificity is underestimated. The predictive values in this scenario are unbiased.

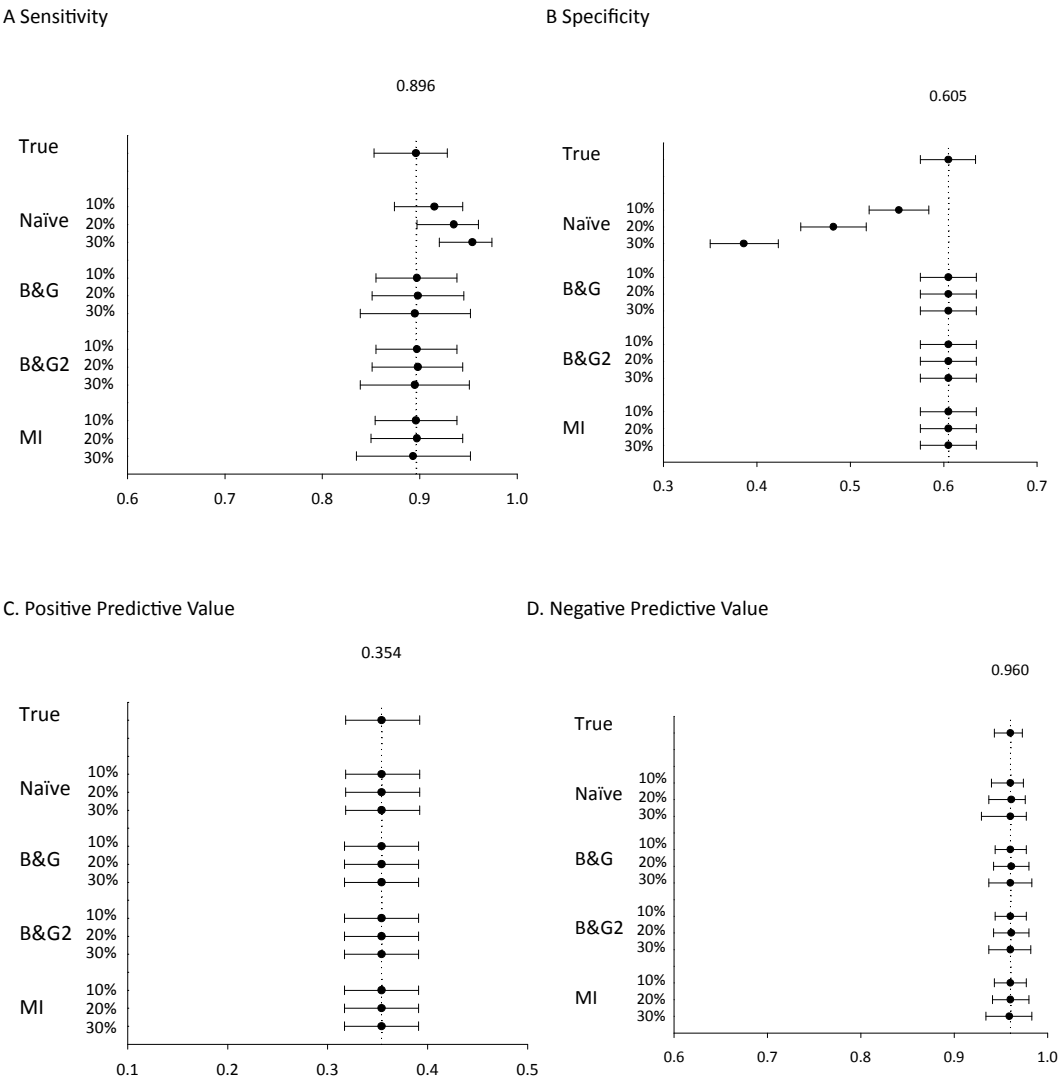

Figure 2. Naive and Corrected Accuracy Measures after Introducing Partial Verification Depending only on the D-dimer Negative Test Result (Partial Verification pattern A).

In this first pattern of missing outcomes the positive predictive values are the same for all correction methods. The positive predictive value is calculated using only the D-dimer positive test results for the diseased and non-diseased. If only the outcomes with a negative test result for the D-dimer test are set to missing, this will not affect the positive predictive values.

For the two correction methods the point estimates of the accuracy measures fluctuate around the true values, even if the proportion of partial verification increases. The 95% confidence intervals, however, become wider with more missing outcomes.

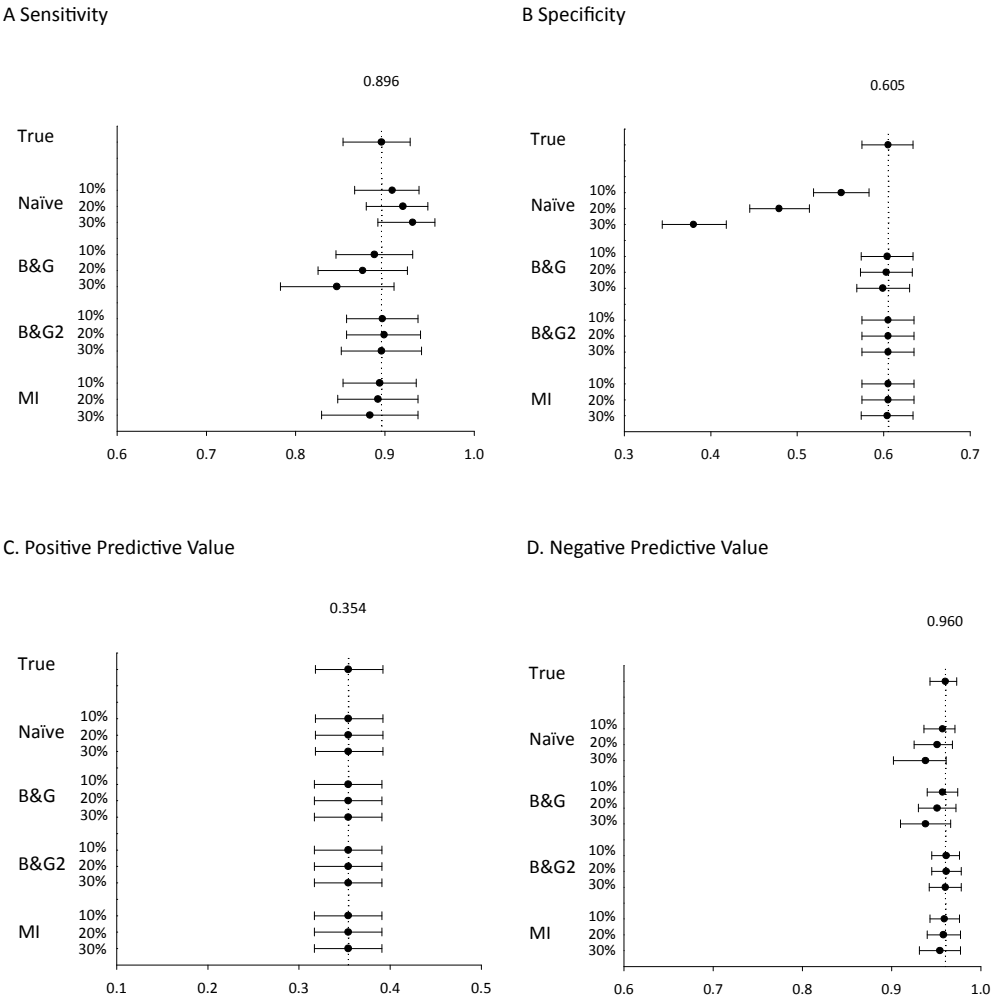

Figure 3. Corrected Accuracy Measures after Introducing Partial Verification Depending on both the Negative D-dimer Test Result as the Negative Calf Circumference Test Result. (Partial Verification pattern B)

With nonreferral based on negative D-dimer and negative Calf Difference test results (Figure 3), an increasing deviation from the true values for sensitivity and specificity is seen with complete case analysis. Again the values for sensitivity are increasingly overestimated whereas the values for specificity are underestimated. In this scenario the negative predictive values are also biased.

The B&G method using only the D-dimer (B&G) hardly reduces this bias and shows results similar to the complete case analysis.

The B&G method using the two relevant predictors (B&G2) and MI can both repair the bias as their point estimates of the accuracy measures fluctuate around the true values, even if the proportion of nonreferral increases. The 95% confidence intervals, however, become wider with an increasing number of missing outcomes, especially in the case of MI.

Because only the reference standard outcome in patients with a negative d-dimer test are set to missing, positive predictive values are again not affected.

In the third scenario, with non-referral based on negative vein distension and no previous surgery (Figure 4), an increasing deviation from the true values for sensitivity and specificity is seen with complete case analysis. The values for sensitivity are overestimated and the values of specificity stay underestimated. The positive predictive values are also underestimated whereas the negative predictive values are estimated correctly.

The B&G method using only the D-dimer test (B&G) partly corrects the bias, because of the correlation between recent surgery/vein distension and the D-dimer test result. For the B&G method using two strong predictors (B&G2) and MI, the point estimates of the accuracy measures fluctuate around the true values, even if the proportion of non-referral increases. The 95% confidence intervals however become wider with an increasing number of missing outcomes.
Because the simulation of missing outcomes was in this case independent of the index test under study, also subjects with a D-dimer positive test results had a chance not to be referred to the reference test. Thus, in this scenario the positive predictive values were also affected.

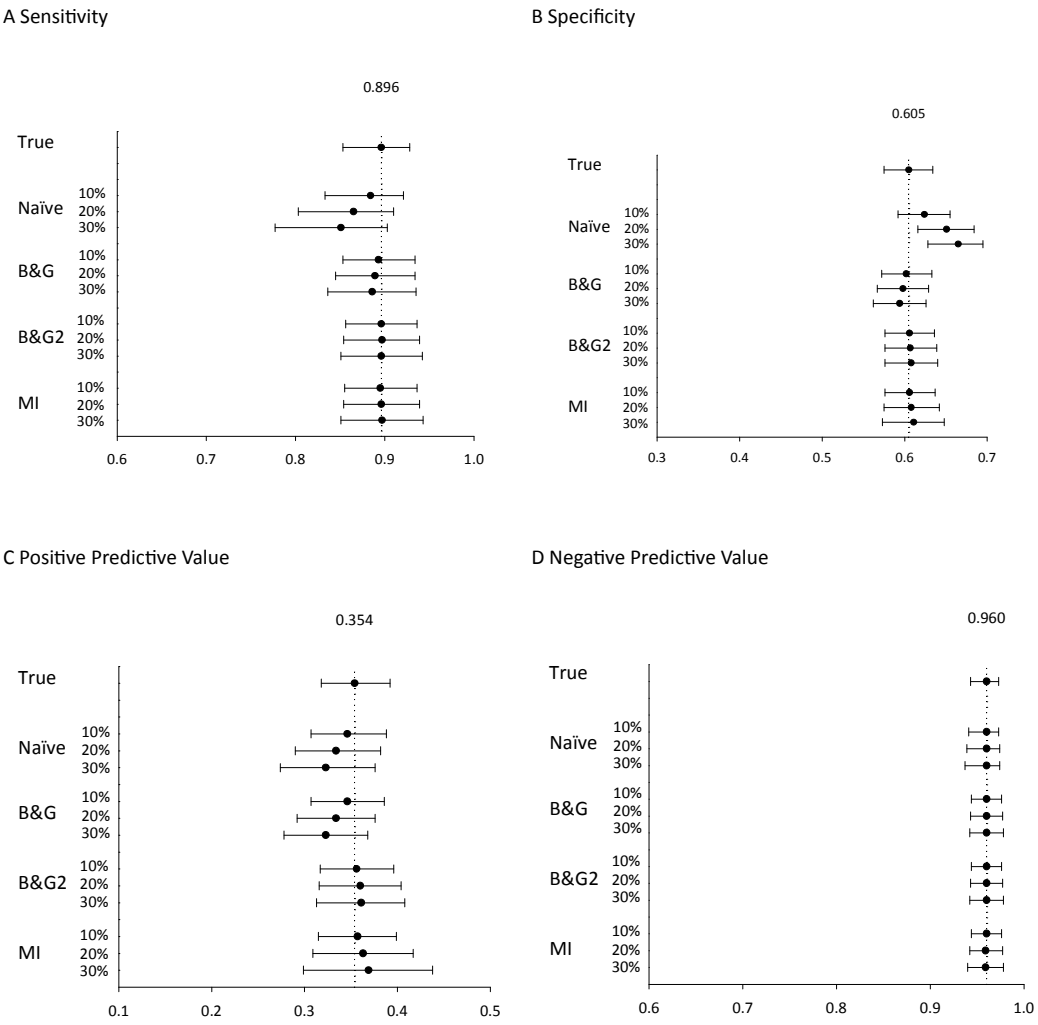

Figure 4. Corrected Accuracy Measures after Introducing Partial Verification Depending on the Absence of Vein Distension in the Legs and the Absence of Recent Surgery (Partial Verification pattern C).

In the figures of all three scenarios, the B&G2 method seems to give slightly more accurate 95% confidence intervals than MI. To further verify this, we calculated the mean-square error (MSE) for every scenario of missing reference data (Table 3).

The MSEs do indeed show that overall the B&G method performs slightly better than MI. The reason is that the B&G method behaves as a maximum likelihood approach for the missing reference data problem. If we would take an unlimited amount of imputation rounds, MI would approach the results of the B&G method even further, and become identical eventually.

Table 3. Mean squared errors for every correction method in all 9 scenarios of missing reference data

	Pattern A; 10% missing				Pattern B; 10% missing				Pattern C; 10% missing			
	Sens	Spec	PPV	NPV	Sens	Spec	PPV	NPV	Sens	Spec	PPV	NPV
naive	0.00045	0.00283	0	0.00002	0.00014	0.00295	0	0.00002	0.00020	0.00039	0.00015	0.00001
B&G1	0.00008	0.00000	0	0.00002	0.00011	0.00000	0	0.00002	0.00005	0.00002	0.00015	0.00001
B&G2	0.00008	0.00000	0	0.00001	0.00004	0.00000	0	0.00001	0.00004	0.00001	0.00007	0.00001
MI	0.00010	0.00000	0	0.00002	0.00005	0.00000	0	0.00001	0.00005	0.00001	0.00009	0.00001

	Pattern A; 20% missing				Pattern B; 20% missing				Pattern C; 20% missing			
	Sens	Spec	PPV	NPV	Sens	Spec	PPV	NPV	Sens	Spec	PPV	NPV
naive	0.00147	0.01525	0	0.00004	0.00055	0.01587	0	0.00012	0.00130	0.00202	0.00064	0.00002
B&G1	0.00019	0.00000	0	0.00004	0.00062	0.00001	0	0.00012	0.00019	0.00009	0.00064	0.00002
B&G2	0.00019	0.00000	0	0.00003	0.00018	0.00000	0	0.00003	0.00010	0.00002	0.00016	0.00002
MI	0.00026	0.00000	0	0.00005	0.00018	0.00000	0	0.00003	0.00011	0.00004	0.00034	0.00002

	Pattern A; 30% missing				Pattern B; 30% missing				Pattern C; 30% missing			
	Sens	Spec	PPV	NPV	Sens	Spec	PPV	NPV	Sens	Spec	PPV	NPV
naive	0.00339	0.04816	0	0.00007	0.00118	0.05059	0	0.00054	0.00233	0.00377	0.00098	0.00002
B&G1	0.00040	0.00000	0	0.00007	0.00260	0.00003	0	0.00054	0.00027	0.00013	0.00098	0.00002
B&G2	0.00038	0.00000	0	0.00007	0.00054	0.00001	0	0.00010	0.00013	0.00002	0.00013	0.00002
MI	0.00055	0.00001	0	0.00011	0.00038	0.00000	0	0.00008	0.00014	0.00011	0.00085	0.00003

Sens=Sensitivity; Spec=Specificity; PPV=Positive Predictive Value; NPV=Negative Predictive Value
Pattern A: non-referral based on negative D-dimer test result
Pattern B: non-referral based on negative D-dimer test result
Pattern C: non-referral based on negative D-dimer test result

Discussion

We studied the correction for partial verification bias with two different correction methods: the B&G method and MI. To study how these correction methods behave in different patterns of partial verification, we studied three different mechanisms leading to partial outcome verification.

Even if partial verification is only based on the index test result, a complete case analysis gives biased results, especially for sensitivity and specificity. This occurs because the pattern in which the missing reference test results were created was clearly a selection of the total study group and not a random sample. The amount of bias increases as the percentages of missing outcomes increases.

The predictive values are in these simulations hardly affected, because the missing reference test results were simulated under the assumption that within the strata of the dependent variables the distribution is random. In different scenarios, of course, which are not unlikely in reality, a complete case analysis can also bias the predictive values. The effects of verification bias on the positive and negative predictive values are well described in literature.[16,17]

The B&G and MI methods are both designed to correct for partial verification bias by using available information. In straightforward situations of missing reference standard data they will therefore lead to similar results. This is reflected in the results of partial verification patterns A and B.

When the mechanism responsible for the missing reference standard results is less straightforward or unknown using MI is probably a safer approach. This method uses all available information of the remaining data, to give a reliable estimation of the missing outcomes. In theory this could also be achieved by using all variables available in the B&G method. In practice, however, this is laborious, because it is not readily available in present statistical software. Furthermore, the B&G method is less straightforward when variables that produce the partial verification are categorical or continuous. MI can easily incorporate categorical and continuous variables in its model using standard available software. Even if the relations between the variables are not linear. This can be solved using log transformations or spline functions.

Although we illustrated the methods using only one practical example of partial verification bias in different patterns of missing reference data, there are no reasons to believe that the main conclusions will not hold in other settings. However, both the B&G method and MI require the missing at random (MAR) assumption. This means that given the observed

data, the reason for missingness does not depend on other unobserved data. In settings where this MAR assumption is violated, methods for correcting non-ignorable verification bias should be considered.[18]

Partial verification by design can be a very efficient data collection strategy. In that case the pattern of missing reference data will be known and accuracy measures can easily be correctly adjusted using either the B&G method, or MI. If not defined by design, partial verification should be avoided, as it can seriously bias the results. There are however situations where the mechanism of missing reference data is not known and partial verification cannot be avoided. In these situations, we recommend to use MI methods to correct. These methods are readily available in statistical software, more flexible than the B&G method, and give reliable estimates of the missing reference data.

References

1. Knottnerus, J. A. and C. van Weel. General introduction: evaluation of diagnostic procedures. In The evidence base of clinical diagnosis. London: BMJ Books, 2002.
2. Whiting, P., A. W. Rutjes, J. B. Reitsma et al., "Sources of variation and bias in studies of diagnostic accuracy: a systematic review." Ann.Intern.Med 140 (2004): 189-202.
3. Reitsma, J. B., A. W. Rutjes, K. S. Khan et al., "A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard." J.Clin.Epidemiol. 62 (2009): 797-806.
4. Keith, C. J., K. A. Miles, M. R. Griffiths et al., "Solitary pulmonary nodules: accuracy and cost-effectiveness of sodium iodide FDG-PET using Australian data." Eur.J.Nucl.Med. Mol.Imaging 29 (2002): 1016-1023.
5. Irwig, L., P. P. Glasziou, G. Berry et al., "Efficient study designs to assess the accuracy of screening tests." Am.J.Epidemiol. 140 (1994): 759-769.
6. Elhendy, A., R. T. van Domburg, D. Poldermans et al., "Safety and feasibility of dobutamine-atropine stress echocardiography for the diagnosis of coronary artery disease in diabetic patients unable to perform an exercise stress test." Diabetes Care 21 (1998): 1797-1802.
7. Begg, C. B. and R. A. Greenes, "Assessment of diagnostic tests when disease verification is subject to selection bias." Biometrics 39 (1983): 207-215.
8. Harel, O. and X. H. Zhou, "Multiple imputation for correcting verification bias." Stat Med 25 (2006): 3769-3786.
9. Hanley, J. A. and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve." Radiology 143 (1982): 29-36.
10. de Groot, J. A., K. J. Janssen, A. H. Zwinderman et al., "Multiple imputation to correct for partial verification bias revisited." Stat Med 27 (2008): 5880-5889.

11. Oudega, R., K. G. Moons, and A. W. Hoes, "Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing." Thromb.Haemost. 94 (2005): 200-205.
12. Altman, D. G. and J. M. Bland, "Diagnostic tests. 1: Sensitivity and specificity." BMJ 308 (1994): 1552.
13. Newcombe, R. G., "Two-sided confidence intervals for the single proportion: comparison of seven methods." Stat.Med. 17 (1998): 857-872.
14. Wilson, EB., "Probable inference, the law of succession, and statistical inference." Journal of the American Statistical Association 22 (1927): 209-212.
15. Rubin, D. B. Multiple imputation for non response in surveys. New York: Wiley, 1987.
16. Zhou XH. Effect of verification bias on positive and negative predictive values. Stat Med. 1994; 13:1737-1745.
17. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford University Press; 2003:176.
18. Kosinski AS, Barnhart HX. Accounting for non-ignorable verification bias in assessment of diagnostic tests. Biometrics. 2003;59:163-171

Chapter 4

Differential verification bias

Chapter 4.1

Adjusting for differential verification bias in diagnostic accuracy studies: A Bayesian approach

Joris A.H. de Groot
Nandini Dendukuri
Kristel J.M. Janssen
Patrick M.M. Bossuyt
Karel G.M. Moons

Epidemiology 2011; 22(2):234-241

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.

Introduction

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]

A Frequently used but conceptually wrong method to handle differential verification

(i)	Verification with preferred reference test			(ii)	Verification with alternative reference test			(iii)	Verification with either reference test	
	R+	R-			S+	S-			**?** +	**?** -
T+	a	b	**+**	T+	-	-	**=**	T+	a	b
T-	-	-		T-	c	d		T-	c	d

T=(index) test under study; R= preferred reference standard; S= alternative reference standard

B Numerical example Appelboam et al: total number of patients (adults only) that underwent Radiology or Follow Up as verification for their Elbow Extension Test result

(i)	Verification with Radiography			(ii)	Verification with Follow Up			(iii)	Verification with either reference test	
	Fracture	no fracture			Fracture	no fracture			**?** +	**?** -
EET+	311	336	**+**	EET+	NA*	NA*	**=**	EET+	311	336
EET-	*2 †*	*56 †*		EET-	3	250		EET-	5	306

† Due to protocol violations a random sample of adult patients who tested negative on EET received radiography

* NA = Results not available and can not be estimated, because all index test positives underwent only the preferred reference test (radiography) and not the alternative reference method (clinical follow-up).

Figure 1. Ambiguous Method to Address Differential Verification

Figure 1A shows the analysis commonly applied in studies of a dichotomous index test in which differential disease verification is used. Results of the two 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 (Figure 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 Figure 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 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, measure 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. 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=(index)Test; R=preferred reference test; S=Alternative reference test; vR_1=proportion of index test positives verified using the preferred reference test; vR_0=proportion of index test negatives verified using the preferred reference test

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:

$$P(T=1)=P(D=1)P(T=1|D=1)+P(D=0)P(T=1|D=0)$$

$$=\pi sT+(1-\pi)(1-cT)$$

The likelihood contribution of the first stage is the probability of observing T_1 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:

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

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:

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

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 condition 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 (α = 1, β = 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 2,000 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).

Table 2. True Parameter Values for the Two Simulated Designs with Different Strategies of Differential Verification

	Design 1			Design 2	
	Sensitivity with respect to D*	Specificity with respect to D*		Sensitivity with respect to D*	Specificity with respect to D*
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**	

* True sensitivities and specificities with respect to the disease status (D).
D= True Disease status; T= Index Test; R=preferred Reference test; S=Alternative reference test; π= disease prevalence; vR_1=proportion of index test positives verified using the preferred reference test; vR_0=proportion of index test negatives verified using the preferred reference test.

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]

Table 3. Simulated data sets from two differential verification designs

a. Data from Differential Verification Design 1					b. Data from Differential Verification Design 2				
Stage of Study			T=1	T=0	Stage of Study			T=1	T=0
Stage I			304	496	Stage I			304	496
Stage II	Probability of Verification on R		1	0	Stage II	Probability of Verification on R		0.7	0.3
	Probability of Verification on S		0	1		Probability of Verification on S		0.3	0.7
Stage III		R=1	116	0	Stage III		R=1	81	20
		R=0	188	0			R=0	132	128
		S=1	0	108			S=1	37	76
		S=0	0	388			S=0	54	272

T=(index)Test; R=preferred reference test; S=Alternative reference test

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.

Analysis of the two 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. 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.

Table 4. Summary of Results of Analyses of Simulated Data from Differential Verification Design 1

	Sensitivity (CI)	Specificity (CI)	PPV (CI)	NPV (CI)
Accuracy measures T wrt 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 wrt 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 (seperate tables)	NA	NA	0.38 (0.33,0.44)	NA
Accuracy measures T wrt 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 (seperate tables)	NA	NA	NA	0.78 (0.74,0.82)
*Accuracy measures T wrt **?***				
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)

*Posterior median and 95% credible intervals; T=(index)Test; D=(latent) Disease status; R=(preferred) Reference test; S=Alternative reference test; PPV=Positive Predictive Value; NPV=Negative Predictive Value; wrt=with respect to; CI: Credible interval

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 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 vs. R and T vs. S.

Table 5. Summary of Results of Different Analyses Based on Simulated Data from Differential Verification Design 2

	Sensitivity (CI)	Specificity (CI)	PPV (CI)	NPV (CI)
Accuracy measures T wrt 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 wrt 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 (seperate 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 wrt 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 (seperate 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 wrt **?***				
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)

T=(index)Test; D=(latent) Disease status; R=(preferred) Reference test; S=Alternative reference test; PPV=Positive Predictive Value; NPV=Negative Predictive Value; wrt=with respect to

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 (i.e., 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.

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		
	Range (%)	Beta coefficients		Range (%)	Beta coefficients	
		α	β		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

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.”

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 proportion of verified subjects (vR1 and vR0) 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.

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

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

NPV = Negative Predictive Value; PPV = Positive Predictive Value; CI = 95% Credible Interval

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 (Figure 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.

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(6):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(3):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(8):797-806.
5. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics.* 1983;39(1):207-215.
6. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. *Stat Methods Med Res.* 1998;7(4):337-353.
7. Harel O, Zhou XH. Multiple imputation for correcting verification bias. *Stat Med.* 2006;25(22):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(28):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(3):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.

11. 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.
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):iii, ix-51.
13. Rutjes AW, Reitsma JB, Irwig LM, Bossuyt PM. Sources of bias and variation in diagnostic accuracy studies. In: AWS Rutjes, editor. Partial and differential verification in diagnostic accuracy studies. Amsterdam: Rutjes; 2005. 31-44.
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(4):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(12):599-610.
17. Plassman BL, Khachaturiane AS, Townsend JJ et al. Comparison of clinical and neuropathologic diseases of alzheimers disease in 3 epidemiologic samples. *Alzheimer's and Dementia.* 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(2):162-177.
19. Buzoianu M, Kadane JB. Adjusting for verification bias in diagnostic test evaluation: a Bayesian approach. *Stat Med.* 2008;27(13):2453-2473.
20. Martinez E.Z., Achcar J.A., Louzada-Neto F. Estimators of sensitivity and specificity in the presence of verification bias: A Bayesian approach. *Computational Statistics & Data Analysis.* 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(3):263-272.
22. Kaldor J, Clayton D. Latent class analysis in chronic disease epidemiology. *Stat Med.* 1985;4(3):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(9):923-937.
24. Gelman A, Rubin DB. Inference from Iterative Simulation using Multiple Sequences. *Statitical Science.* 1992;7(4):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(5):560-566.

26. Pudas T, Hurme T, Mattila K, Svedstrom E. Magnetic resonance imaging in pediatric elbow fractures. *Acta Radiol.* 2005;46(6):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(7):633-636.
28. Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. *Clin Chem.* 2004;50(3):473-476.
29. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics.* 2001;57(1):158-167.
30. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics.* 1985;41(4):959-968.

Chapter 5

Diagnostic accuracy studies without a single, acceptable reference standard: a case study

Joris A.H. de Groot
Patrick M.M. Bossuyt
Johannes B. Reitsma
Nandini Dendukuri
Kristel J.M. Janssen
Karel G.M. Moons

Introduction

In studies of diagnostic accuracy, the results of one or more tests under evaluation are compared with the results of a reference standard, both measured in subjects who are suspected of having the condition of interest. This reference standard is considered to be the best available method for establishing the presence or absence of the condition of interest. [1-4] However, for several target conditions there is no generally accepted reference standard available. Heart failure, depression and migraine are common clinical examples of conditions without a proper reference standard.

The validity of diagnostic accuracy studies directly depends on the quality of the reference standard procedure. If making accurate clinical diagnosis is challenging because of the absence of a single proper reference standard, clinicians often determine the presence or absence of the target condition in each patient based on multiple sources of information. These sources can include general patient characteristics, signs and symptoms from history and physical examination, and other rest results.[4]

It is well known that there can be substantial variation in the use of tests and subsequent treatment decisions between physicians when dealing with patients presenting with similar symptoms.[5] With the limitations of individual clinician diagnoses in mind, research studies often use the consensus coming from a panel of several clinical experts. Such a panel may achieve greater diagnostic reliability, accuracy, and certainty than an individual expert.[6] Although this seems a clinically appealing approach, it is questionable whether this is the most optimal solutions, as such classifications can vary with the quality and background of experts, presence of powerful personalities, overestimation of certain test results etc.[4]

Similar to clinicians combining several diagnostic test results to determine the presence or the absence of the target condition in a patient, one can also use mathematical modelling to determine disease status. This is called a latent class analysis and is based on the idea that observed variables are jointly determined by an underlying, unobserved (latent) target condition. In diagnostic accuracy studies the true disease status can be considered as a dichotomous latent variable with two categories, 'disease present' and 'disease absent'. Within a group of individuals with unknown disease status, for whom at least three independent diagnostic test results are available, both disease prevalence and sensitivity and specificity of all tests can be estimated from the pattern of diagnostic tests results using latent class analysis.[7-9]

A latent class analysis is a statistically sound and flexible approach. A major potential benefit is that it is more objective than a panel diagnosis as it is examining the strength of statistical

relationships among variables without preference for a specific test.[1] However, since disease is defined in statistical terms clinicians can feel uncomfortable what the results stand for.

In this study we introduce the idea to directly compare the two different verification strategies and discuss the possible origin of observes differences. We illustrate the application of both methods using the data from a heart failure study. The results of this case study may generate further ideas for analysing and designing studies to critically compare different solutions for diagnostic accuracy studies with verification problems.

Methods

To illustrate the use of consensus panel diagnosis and latent class analysis in a diagnostic accuracy study without a proper reference standard, we use data from a study in patients suspected for heart failure.

Table 1. Characteristics of patients suspected for heart failure. Heart failure as classified as present or absent based on a consensus panel diagnosis.

	all		heart failure (based on panel diagnosis)			
			yes		no	
	n=721		n=207		n=514	
	n	%	n	%	n	%
female gender	466	64.6	125	60.4	341	66.3
mean age in years (sd)	70.7	(11.8)	75.5	9.7	68.8	12.1
medical history						
- *MI, CABG or PTCA*	48	6.7	26	12.6	22	4.3
physical examination						
- mean systolic blood pressure (sd)	156.0	(26.7)	153.8	30.0	157.0	25.3
- mean diastolic blood pressure (sd)	87.0	(13.2)	86.5	15.0	87.1	12.4
- mean pulse (sd)	77.2	(14.8)	82.5	16.8	75.1	13.3
- wheezing or rhonchi	58	8.0	15	7.3	43	8.4
- irregular pulse	72	10.0	51	24.6	21	4.1
- *displaced apex beat*	70	9.7	51	24.6	19	2.6
- heart murmer suggesting mitral regurgitation	77	10.7	46	22.2	31	6.0
- elevated jugular venous pressure	56	7.8	37	17.9	19	3.7
ECG						
- 'abnormal' ECG	536	74.3	192	92.8	344	66.9
Spirometry						
- Predicted percentage vital capacity (sd)	97.4	(20.3)	89.3	21.9	100.5	18.7
BNP						
- mean log NT-proBNP (sd)	3.40	(1.7)	4.85	1.86	2.84	1.28

sd= standard deviation; MI=Myocardial Infarction; CABG=Coronary Artery Bypass Graft; PTCA=Percutaneous Transluminal Coronary Angioplasty; ECG=ElectroCardioGram; BNP= B-type Natriuretic Peptide; NT-proBNP= N-terminal proBNP.

In this study, 721 patients suspected for heart failure by their general practitioner were included. All underwent a standardized diagnostic work-up. A selection of the most important baseline characteristics is shown in Table 1.

To avoid the problem of missing values in the panel and latent class analysis, we excluded those patients with missing values on one of the three variables used in the latent class model (n=67) as this will not harm the illustrative goals of this example.

Panel Diagnosis

Since there is no universally accepted single reference standard available for heart failure, the researchers, in accordance with earlier studies,[10-13] used a consensus panel diagnosis to decide on the presence or absence of the target condition. Each panel consisted of a cardiologist, a pulmonologist, and the principal investigator, who is a general practitioner working in an outpatient heart failure clinic. In all there were 12 meetings in which the cardiologist or pulmonologist could vary, but the principal investigator was present in every panel. A panel evaluated all the available diagnostic test results including a 6 months follow-up period from each patient and classified each patient as heart failure present or absent using the criteria outlined by the European Society of Cardiology.[14]

Latent Class model

To make a comparison between the results of a latent class model and the consensus panel decision, we set up a latent class model using three relevant clinical patient characteristics:

- Medical history: whether or not the patient had either a previous myocardial infarction (MI) or underwent Coronary Artery Bypass Grafting (CABG) or Percutaneous Transluminal Coronary Angioplasty (PTCA);
- Physical examination: whether or not there was a displaced apex beat present;
- Extra testing: whether or not the NT-proBNP measurement was abnormal.

In order to compose a simple latent class model with solely dichotomous tests, we dichotomized the NTproBNP data using age-specific standard reference values for NT-proBNP in adults (Table 2; source: Labo Medical Analysis).

Table 2. Normal values for NT-proBNP in adults by age (source: Labo Medical Analysis)

Age (years)	Males	Females
<50	< 88 pg/mL	< 153 pg/mL
50-65	< 227 pg/mL	< 334 pg/mL

To estimate the parameters of our latent model we used the BayesLatentClassModels (BCLM) program.[15] BCLM is a specialised software package that can fit latent class models to estimate the properties of each diagnostic tests, i.e. sensitivity and specificity, along with disease prevalence. It uses a Bayesian approach that allows prior information on the prevalence, sensitivities or specificities to be incorporated in the analysis. This can be helpful when the problem is non-identifiable.[8;16;17] For more details regarding the statistical methods implemented by BLCM we refer to the literature.[8;18]
We examined three different outcome measures to compare the two verification strategies.

The first outcome we compared is the estimate of the prevalence of heart failure coming from the consensus panel and the latent class model.

The second outcome is the estimates of accuracy from individual tests. Since the three tests being used in the latent class model - previous MI, CABG or PTCA, Displaced Apex Beat, positive NT-proBNP result - are also used by the panel to predict heart failure, we can calculate the sensitivity and specificity of all three tests with regard to the panel diagnosis and to the latent class model.

Perhaps the most clinically relevant comparison between the consensus panel and the latent class approach is the difference in classification of patients on an individual level i.e. in how many patients do the two methods disagree whether heart failure is present or absent. To make such a comparison we need to define an explicit threshold for the latent class results. Where the consensus panel simply decides on heart failure present or absent, the latent class model generates probability estimates of heart failure being present given a specific pattern of test results. After setting a threshold for the probability above which we assume that heart failure is present, we can classify each patient as having heart failure or not. Here, we applied a threshold of 0.5 on the probability of having heart failure being present. This means that every patient with a combination of test results corresponding to a probability of more than 0.5 in the latent class model will be considered as having heart failure. Every patient with a probability below 0.5 will be classified as a patient without heart failure. We then compose a two by two table with the agreements and disagreements in final diagnoses coming from consensus panel and latent class approach.

Results

Prevalence

A final diagnosis of heart failure was made by the consensus panel in 207 patients corresponding to a prevalence of 0.29. The estimate of prevalence based on the latent class model was very similar with a value of 0.28 (Table3).

Table 3. Univariable estimates (95% confidence intervals) of accuracy measures of the 3 different diagnostic tests with respect to both the panel and the latent class diagnosis.

	Accuracy based on panel diagnosis			Accuracy based on latent class model		
Accuracy measures	MI/CABG/ PTCA	Displaced Apex Beat	Abnormal NT-proBNP	MI/CABG/ PCTA	Displaced Apex Beat	Abnormal NT-proBNP
Sensitivity	0.13 (0.09-0.18)	0.25 (0.19-0.31)	0.37 (0.30-0.44)	0.13 (0.07-0.22)	0.30 (0.21-0.56)	0.44 (0.30-0.82)
Specificity	0.96 (0.94-0.97)	0.96 (0.94-0.98)	0.96 (0.94-0.97)	0.96 (0.93-1.00)	0.99 (0.94-1.00)	0.99 (0.92-1.00)
Prevalence heart failure		0.29 (0.25-0.32)			0.28 (0.10-0.42)	

MI=Myocardial Infarction; CABG=Coronary Artery Bypass Graft; PTCA=Percutaneous Transluminal Coronary Angioplasty; BNP= B-type Natriuretic Peptide; NT-proBNP= N-terminal proBNP.

Univariable estimates of test accuracy

Similar to prevalence, the point estimates of the sensitivities and specificities of both methods were highly comparable. The 95% credible intervals using the latent class model are wider than the 95% confidence intervals based on the panel diagnosis approach (Table 3). One reason is that the confidence interval accompanying the panel diagnosis only represents binomial uncertainty, whereas the credible interval represents binomial uncertainty and additional uncertainty around the parameters in the model.

Panel versus latent class diagnosis at the patient level

The 2x2 table in Table 4 compares the final classification in each patient from consensus panel and the latent class model. In 522 patients (79.8%) there is agreement in the diagnoses of patients from he consensus panel and latent class analysis. However, 96 patients that are diagnosed with heart failure by the consensus panel were not classified as such by the latent class model, and 36 patients diagnosed with heart failure by the latent class analysis did not receive this diagnosis by the consensus panel.

Table 4. Agreements and disagreement in final diagnoses coming from consensus panel and latent class analysis

	Latent class analysis	
	Heart failure present	Heart failure absent
Consensus panel		
Heart failure present	88	96
Heart failure absent	36	434

Discussion

In this case study we introduced the idea of a head-tot-head comparison between two different methods to deal with diagnostic accuracy studies in the absence of a proper reference standard. The choice between a consensus panel or a latent class model is a difficult one because both approaches have their merits and pitfalls. Whereas clinical researchers tend to go for the more clinically appealing approach of a consensus panel, statisticians might opt for the more objective and therefore less time or place dependent latent class analysis. More extensive clinical or simulation studies comparing these two verification strategies will enable researchers to better understand when the performances of both methods will differ and when differences will be small. Such studies will enable us to make better methodological choices when confronted with diagnostic research in the absence of a proper reference standard. This study explored how such a study might look like.

When we compared the two methods head-to-head in our heart failure example, results with respect to estimates of accuracy and prevalence were rather similar. However, when comparing the final classification on an individual level differences occurred in 20.2% of the patients. This at least demonstrates that these methods cannot be used interchangeably. A universal problem in such comparison studies will be how to decide which method is correct if the two approaches disagree in their classification. Ideally, a third external method exists which can act as a fair referee. More extensive follow-up with additional testing could be an option.

There are, of course, a few drawbacks to this preliminary case study. In the heart failure study we used, no disease probabilities have been generated by the consensus panel. The experts only decided on disease presence and absence. Disease probabilities could have provided further insight whether the differences in classification indeed occurred in "difficult" patients and whether these differences in part can be explained by the use of a different threshold to define disease. In future studies, these disease probabilities generated by the consensus will be crucial to make an in depth comparison between the two methods. Also, in future research, the latent class method should be further extended by modelling conditional dependence between variables or perhaps incorporating more variables.

It remains unclear whether or when either a consensus panel diagnosis or a latent class model should be used in diagnostic accuracy studies without an available reference standard providing acceptable verification. In some situations, it might even be preferable to acknowledge the fact that a standard diagnostic accuracy studies can or should not be performed and another type of evaluation studies, clinical test validation, should be considered.[1;4]

In our opinion, further research thoroughly examining the similarities and differences between a consensus panel diagnosis and latent class analyses need to be done to better understand the performances of both methods in different clinical settings. This will enable us to make better methodological choices when confronted with diagnostic research in the absence of a proper reference standard.

Reference List

1. 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(8):797-806.
2. de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N., Janssen KJ et al. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ. In press 2011.
3. Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic procedures. The evidence base of clinical diagnosis. London: BMJ Books; 2002. 1-18.
4. 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):iii, ix-51.
5. Elmore JG, Feinstein AR. A bibliography of publications on observer variability (final installment). J Clin Epidemiol 1992; 45(6):567-580.
6. Gabel MJ, Foster NL, Heidebrink JL, Higdon R, Aizenstein HJ, Arnold SE et al. Validation of consensus panel diagnosis in dementia. Arch Neurol 2010; 67(12):1506-1512.
7. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. Biostatistics 2007; 8(2):474-484.
8. 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(3):263-272.
9. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 2001; 57(1):158-167.
10. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003; 138(1):40-44.
11. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988; 240(4857):1285-1293.
12. Rutten FH, Moons KG, Cramer MJ, Grobbee DE, Zuithoff NP, Lammers JW et al. Recognising heart failure in elderly patients with stable chronic obstructive pulmonary disease in primary care: cross sectional diagnostic study. BMJ 2005; 331(7529):1379.

13. Cowie MR, Wood DA, Coats AJ, Thompson SG, Poole-Wilson PA, Suresh V et al. Incidence and aetiology of heart failure; a population-based study. Eur Heart J 1999; 20(6):421-428.
14. Hobbs FD, Jones MI, Allan TF, Wilson S, Tobias R. European survey of primary care physician perceptions on heart failure diagnosis and management (Euro-HF). Eur Heart J 2000; 21(22):1877-1887.
15. BayesianLatentClassModels: A program for estimating diagnostic test properties and disease prevalence [2007.
16. de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Bossuyt PM, Moons KG. Adjusting for Differential-verification Bias in Diagnostic-accuracy Studies: A Bayesian Approach. Epidemiology 2011; 22(2):234-241.
17. Dendukuri N, Belisle P, Joseph L. Bayesian sample size for diagnostic test studies in the absence of a gold standard: Comparing identifiable with non-identifiable models. Stat Med 2010; 29(26):2688-2697.
18. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Stat Med 2009; 28(3):441-461.

Chapter 6

Adjusting for partial verification bias in meta-analyses of diagnostic accuracy studies

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

Provisionally accepted (AJE)

Abstract

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. For a variety of practical and ethical reasons, sometimes only a proportion of patients receives the reference standard, which can bias the accuracy estimates. Numerous methods have been described for correcting this partial verification bias or work-up bias in individual studies. We describe a Bayesian method to obtain adjusted results from a diagnostic meta-analysis when partial verification or work-up 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 our method will be compared to the existing methods for dealing with verification bias in diagnostic meta-analyses. Empirical data from a systematic review of studies of the accuracy of the immunohistochemistry test for diagnosis of HER-2 status in breast cancer patients will be used for illustration.

Introduction

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 relationship 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. Primary studies in a diagnostic meta-analysis may, therefore, 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, work-up or verification bias.[8;12-14] A classical scenario in which this bias arises is a two-stage design, where all subjects undergo the test under evaluation or index test at Stage I, but only a sample of subjects is selected at Stage II to undergo verification of disease presence by the reference standard. When selection of subjects for the reference standard is not a completely random sample, verification bias will occur. This could happen, for example, when a stratified random sample is drawn at Stage II with the strata being defined by the results of the index test in Stage I. Such a non-random referral pattern may arise due to 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 to address this particular form of work-up 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, when partial verification is present in one or more of the primary studies. Some review authors have acknowledged this bias in their discussions, but not quantified nor corrected for it in the analyses.[17-19] A recent article describes a meta analytical method for the case where primary diagnostic studies have missing data on the reference standard but the authors did not explicitly address the problem as partial verification, work-up or selection bias.[11]

Excluding all primary studies with work-up bias is one simple and frequently applied solution.[2] This method avoids the partial verification bias , but at the expense of reduced precision and a lower generalizability. Alternatively, sensitivity analyses that alternatively include and 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.

We extend existing methods to correct for partial verification models for single diagnostic studies[13] to the meta-analytic setting. In particular, we propose a two-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 two 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

Our method will be illustrated using data from a recently published systematic review on testing for HER2-positive breast cancer, an aggressive form of breast cancer associated with a high mortality rate.[21] The availability of Herceptin, an effective but expensive treatment for HER2 positive breast cancer, has increased the awareness of having adequate accuracy in identifying women who have HER2 receptors and are thus most likely to respond to this therapy. Two tests are commonly used to determine HER2 status: immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH). FISH is believed to be the gold standard test for determining HER2 status. It 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 values 0, 1+, 2+ or 3+. Patients who receive scores of 0 or 1+ are considered to be HER-2 negative, while those with scores of 3+ are considered to be HER-2 positive. Patients with a score of 2+ are considered to have an ambiguous test results. Various studies have recommended that the results of patients with IHC scores of 2+ should be verified with a FISH test.[22] More recent studies[21;23] have

recommended that patients with IHC 3+ scores should also be verified by the FISH test. This implies that patients who receive IHC scores of either 2+ or 3+ are more likely to be verified by FISH in routine clinical practice than those who receive 0 or 1+ scores. 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 around 30%.[21] Thus we would expect that studies with a work-up bias have a proportion of IHC 2+ and 3+ scores greater than 30%.

Table 1. Summary of Studies Included in the Meta-Analysis of the Sensitivity and Specificity of Immunohistochemistry (IHC) With Respect to Fluorescence In Situ Hybridization (FISH) for HER-2 Breast Cancer.

Study	No. of patients	IHC score, % of patients			% of patients with positive FISH result in each IHC score category		
		0 and 1+	2+	3+	0 and 1+	2+	3+
Hoang et al, 2000	100	74.0	2.0	24.0	0.0	0.0	70.8
Kakar et al, 2000	112	70.5	15.2	14.3	1.3	35.2	87.5
Bartlett et al, 2001	210	85.2	10.0	4.8	6.7	90.5	90.0
Tsuda et al, 2001	101	76.3	5.9	17.8	2.6	0.0	83.3
Press et al, 2002	117	74.4	11.1	14.5	14.9	100.0	100.0
Dowsett et al, 2001	426	63.4	12.7	23.9	0.7	48.1	94.1
Ogura et al, 2003	110	71.9	9.1	18.2	3.7	10.0	100.0
Lal et al, 2004	2279	76.0	13.7	10.3	1.9	26.5	89.7
Lottner et al, 2005	215	78.1	11.6	10.2	2.4	72.0	100.0
Loring et al, 2005*	110	56.4	15.5	28.2	0.0	0.0	87.1
Lebeau et al, 2001*	78	56.4	20.5	23.1	0.0	25.0	100.0
McCormick et al, 2002*	198	56.6	22.7	20.7	6.3	42.3	100.0
Roche et al, 2002*	119	16.0	10.1	73.9	0.0	0.0	89.8
Mrozkowiak et al, 2004*	360	2.8	87.5	9.7	0.0	20.3	91.4
Yaziji et al, 2004*	2913	49.0	39.5	11.5	2.8	17.0	91.6
Dolan et al, 2005*	129	17.9	72.1	10.1	0.0	7.5	38.4
Press et al, 2005*	842	54.3	14.7	31.0	4.2	16.9	78.2

IHC= Immunohistochemistry; FISH= Fluorescence In Situ Hybridization
*studies with verification bias

The 17 studies included in this meta-analysis 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+ immunohistochemistry scores.[24-31] In four of these studies it was evident from the methods that the study sample was collected at a centre where patients had been

selectively referred for FISH.[24-26;28] In one study, the study design involved oversampling of patients with IHC 2+ results.[30] We treated an additional three 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%.[27;29;31]
In studies without verification bias, the percentage of patients in the IHC 0 or 1+ categories ranged from 63% to 85% compared to 2.8% to 57% in the studies that were considered to have verification bias.

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

General approach
We assumed that all primary studies followed a two-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 were verified by 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 at Stage II. However, in those studies where there was verification bias, an unknown proportion of patients received the reference standard at Stage II. Studies with verification bias would typically report only results based on the subset of patients verified in Stage II.

The meta-analysis was also carried out in two stages to reflect the two-stage data collection process. In the first stage of the meta-analysis we estimated the probability distribution (i.e. 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(Reference+|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 to implement the model is given in the Appendix. Finally, the pooled sensitivity and specificity of the index test were obtained as functions of the parameters estimated in the two stages of the meta-analysis.

Stage I Distribution of (index) test results
We assumed the index test results are expressed on an ordinal scale while the reference standard results are dichotomous. Let (t_{1j}, t_{2j}, …, t_{lj}) denote the number of subjects in the

jth study with results 1, ..., 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}, …, p_{Ij}) and sample size $n_i = t_{1j} + t_{2j} + \ldots + t_{Ij}$. 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 two cumulative probabilities, $p_{ij} = q_{ij} - q_{i-1j}$ (32). Each q_{ij} can be expressed as a probit (cumulative normal probability) function of a continuous variable a_{ij}, i.e. $q_{ij} = \Phi(a_{ij})$. 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 $N(A_i, \sigma_i)$, ai-1j <= aij <= ai+1j, where A_i denotes the pooled mean value of the a_{ij} 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 $a1_j$ is $-\infty$ and the upper limit of truncation for a_{Ij} is ∞. We used objective N(mean=0, 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 j^{th} study we observe the variables r_{ij}, i=1, … , 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}, i.e. $s_{ij} = \Phi(b_{ij})$. 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} across all studies and τ_i is the between study standard deviation. Once again objective prior distributions 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 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 Monte Carlo Markov Chain (MCMC) methods. For each model described in this paper, five MCMC runs were carried out with different starting values. Convergence of the model was determined using the Gelman-Rubin statistic provided by WinBUGS.[33] 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, 2.5% and 97.5% quantiles).

Estimating the pooled sensitivity and specificity of the index test

Let $P_1 = \Phi(A_1)$, $P_2 = \Phi(A_2) - \Phi(A_1)$, ..., $P_I = 1 - \Phi(A_{I-1})$ denote the pooled estimates across all

studies of the prevalence of each value of the index test. Similarly, let $S_1 = \Phi(B_1)$, $S_2 = \Phi(B_2)$, ..., $S_I = \Phi(B_I)$ 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 cut-off 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 cut-off 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 cut-off for defining verification bias P(IHC=2+ or 3+)>30%. This would imply that the study by Dowsett et al[34] 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 Spiegelhalter et al[35] to assess the sensitivity of our inferences to commonly used low information prior distributions. In addition to the U(0, 100) prior over the standard deviation, we fit the model with an Gamma(0.001, 0.001) prior over the between-study precision and a U(0, 100) prior over the between-study variance.

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

We compared the results of the two-stage model described above to the results obtained when: i) verification bias is ignored, and ii) when studies with verification bias are excluded in total from the analysis.

Results

Distribution of index test and reference standard results

Table 2a shows the posterior estimates (median and 95% credible intervals) derived from Stage I of the model, for the probability of each value of the IHC test. Table 2b shows the posterior estimates (median 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 positive FISH results (especially the 2+ and 3+ categories) indicate that there was a substantial amount of between study variability.

Table 2. Overall Results of the Meta-Analysis for the Percentages of Patients in each IHC Category (Stage I) and the Percentage of Patients With a Positive FISH Test Result in each IHC Category (Stage II).

a. IHC score, probability (Stage I)

	IHC 0 and 1+		IHC 2+		IHC 3+	
Overall	Median	95% CI	Median	95% CI	Median	95% CI
	0.77	0.73, 0.80	0.11	0.05, 0.10	0.13	0.09, 0.18

b. Probability of a positive FISH result in each IHC score category (Stage II)

	IHC 0 and 1+		IHC 2+		IHC 3+	
Overall	Median	95% CI	Median	95% CI	Median	95% CI
	0.03	0.02, 0.04	0.27	0.11, 0.48	0.91	0.85, 0.95

IHC: Immunohistochemistry; CI: Credible Interval

Pooled sensitivities en specificities per correction method

Table 3 lists the pooled estimates of sensitivity (Table 3a) and specificity (Table 3b) 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.

At the cut-off of 2+, we found that the sensitivity obtained with the method that ignored verification bias altogether was higher while specificity was considerably lower compared to the results from the adjusted model. At the cut-off of 3+, the pattern was reversed with the sensitivity being lower than in the adjusted model. This was probably because the IHC 2+ sub-group was more likely to be over-sampled at Stage II than the IHC 3+ sub-group.

These results are similar to what has been found when ignoring verification bias in primary diagnostic studies,[8;12-14] i.e. when oversampling index test positive subjects (IHC>=2+) the sensitivity is overestimated while the specificity is underestimated, while when oversampling index test negative patients (IHC<=2+) the sensitivity is underestimated while the specificity is overestimated.

Table 3. Sensitivities (a.) and Specificities (b.) of the Possible IHC Scores Using our Bayesian Method and Three Other Methods: i) Ignoring Verification Bias in the Analysis, and ii) Excluding Studies With Work-Up Bias in the Analysis.

a. Sensitivity of IHC at Different Cut-Offs

IHC score cut-off	Verification bias corrected using the Bayesian Method		i) Ignoring verification bias		ii) Excluding studies with verification bias	
	Median	95% CI	Median	95% CI	Median	95% CI
≥ 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	

b. Specificity of IHC at Different Cut-Offs

IHC score cut-off	Verification bias corrected using the Bayesian Method		i) Ignoring verification bias		ii) Excluding studies with verification bias	
	Median	95% CI	Median	95% CI	Median	95% CI
≥ 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	

IHC: Immunohistochemistry; CI: Credible Interval

Sensitivity analyses

Sensitivity analyses, altering the cut-off to classify a study as having verification bias, did not have an important impact on the pooled median and 95% credible interval of the sensitivities and specificities (Sensitivity at 2+: 0.88 (0.81, 0.92), Sensitivity at 3+: 0.71 (0.56, 0.82), Specificity at 2+: 0.90 (0.85, 0.93), Specificity at 3+: 0.99 (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

We presented a method to adjust for work-up 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 similar results as the corrected values achieved by our Bayesian model. However, leaving studies completely out of the analysis can in general 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 of the analysis (due to verification bias), the more this will affect both bias and precision. Meta-analyses are in principal 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 the methods section of 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 based on clinical and methodological grounds. Because this is a more subjective assessment it may increase the risk of misclassification of studies and 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 Begg and Greenes method has been thoroughly studied.[13, 36] 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 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.[37, 38] 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 important particularly if additional covariates besides the index test results determine the probability of verification, and also affect 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 a meta-analysis of such studies can seriously harm the 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.

References

1. Leeflang MM et al. Systematic reviews of diagnostic test accuracy. Ann.Intern.Med 2008;149:889-97.
2. Irwig L et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann.Intern. Med. 1994;120:667-76.
3. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865-84.
4. Reitsma JB et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin.Epidemiol. 2005;58:982-90.
5. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR Am J Roentgenol. 2006;187:271-81.
6. Zhou XH, Brizendine EJ, Pritz MB. Methods for combining rates from several studies. Stat Med 1999;18:557-66.
7. Whiting P et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann.Intern.Med 2004;140:189-202.
8. Rutjes AW et al. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174:469-76.
9. Sadatsafavi M 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. 2009.
10. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. J.Clin.Epidemiol. 1999;52:943-51.
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;104:512-23.

12. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926-30.
13. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:207-15.
14. Lijmer JG et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.
15. Harel O, Zhou XH. Multiple imputation for correcting verification bias. Stat Med 2006;25:3769-86.
16. de Groot JA et al. Multiple imputation to correct for partial verification bias revisited. Stat Med 2008;27:5880-9.
17. Met R et al. Diagnostic performance of computed tomography angiography in peripheral arterial disease: a systematic review and meta-analysis. JAMA 2009;301:415-24.
18. Nayak S et al. Meta-analysis: accuracy of quantitative ultrasound for identifying patients with osteoporosis. Ann.Intern.Med 2006;144:832-41.
19. Mijnhout GS et al. Systematic review of the diagnostic accuracy of (18) F-fluorodeoxyglucose positron emission tomography in melanoma patients. Cancer 2001;91:1530-42.
20. Wells PS 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;122:47-53.
21. Dendukuri N et al. Testing for HER2-positive breast cancer: a systematic review and cost-effectiveness analysis. CMAJ. 2007;176:1429-34.
22. Hsi ED, Tubbs RR. Guidelines for HER2 testing in the UK. J.Clin.Pathol. 2004;57:241-2.
23. Elkin EB et al. HER-2 testing and trastuzumab therapy for metastatic breast cancer: a cost-effectiveness analysis. J.Clin.Oncol. 2004;22:854-63.
24. Yaziji H et al. HER-2 testing in breast cancer using parallel tissue-based methods. JAMA 2004;291:1972-7.
25. Roche PC et al. Concordance between local and central laboratory HER2 testing in the breast intergroup trial N9831. J.Natl.Cancer Inst. 2002;94:855-7.
26. Press MF 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;11:6598-607.
27. Lebeau A 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;19:354-63.
28. 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;123:766-70.

29. McCormick SR et al. HER2 assessment by immunohistochemical analysis and fluorescence in situ hybridization: comparison of HercepTest and PathVysion commercial assays. Am.J.Clin.Pathol. 2002;117:935-43.
30. Mrozkowiak A et al. HER2 status in breast cancer determined by IHC and FISH: comparison of the results. Pol.J.Pathol. 2004;55:165-71.
31. Loring P 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;13:194-200.
32. Tosteson AN et al. ROC curve regression analysis: the use of ordinal regression models for diagnostic test assessment. Environ.Health Perspect. 1994;102 Suppl 8:73-8.
33. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science 1992;457-511.
34. Dowsett M 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;199:418-23.
35. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. London: Chapman&Hall/CRC, 1995.
36. de Groot JA et al. Correcting for partial verification bias: a comparison of methods. Ann. Epidemiol. 2011;21:139-48.
37. Lu Y et al. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. Stat.Med. 2010;29:2532-43.
38. 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;141:263-72.

Appendix 1. WinBugs Model for Correcting Verification Bias in a Meta-Analysis

```
model {
# Meta-Analysis Stage I
for (j in 1:9) {
n[j]<-sum(r[j,1:6])

t[j,1:3] ~ dmulti(p[j,1:3],n[j])

p[j,1]<-q[j,1]
p[j,2]<-q[j,2]-q[j,1]
p[j,3]<-1-p[j,1]-p[j,2]

logit(q[j,1])<-a[j,1]
logit(q[j,2])<-a[j,2]

a[j,1] ~ dnorm(A[1],pr.a[1]) I(-5,a[j,2])
a[j,2] ~ dnorm(A[2],pr.a[2]) I(a[j,1],5)

}

A[1] ~ dnorm(0,0.0001) I(-5,A[2])
A[2] ~ dnorm(0,0.0001) I(A[1],5)

pr.a[1]<-pow(σ [1],-2)
σ [1] ~ dunif(0,100)
pr.a[2]<-pow(σ [2],-2)
σ[2] ~ dunif(0,100)


# Meta-Analysis Stage II

for (j in 1:17) {

# Total number in each IHC category
t[j,1]<-r[j,1]+r[j,4]
t[j,2]<-r[j,2]+r[j,5]
t[j,3]<-r[j,3]+r[j,6]

r[j,1] ~ dbin(s[j,1],t[j,1])
r[j,2] ~ dbin(s[j,2],t[j,2])
r[j,3] ~ dbin(s[j,3],t[j,3])

logit(s[j,1]) <-b[j,1]
logit(s[j,2]) <-b[j,2]
logit(s[j,3]) <-b[j,3]
```

```
b[j,1] ~ dnorm(B[1],pr.b[1]) I(-5,5)
b[j,2] ~ dnorm(B[2],pr.b[2]) I(-5,5)
b[j,3] ~ dnorm(B[3],pr.b[3]) I(-5,5)

}

B[1] ~ dnorm(0,0.0001) I(-5,5)
B[2] ~ dnorm(0,0.0001) I(-5,5)
B[3] ~ dnorm(0,0.0001) I(-5,5)

pr.b[1]<-pow(τ [1],-2)
τ [1] ~ dunif(0,100)
pr.b[2]<-pow(τ [2],-2)
τ [2] ~ dunif(0,100)
pr.b[3]<-pow(τ [3],-2)
τ [3] ~ dunif(0,100)

# Distribution of IHC scores
qq[1]<-1/(1+exp(-A[1]))
qq[2]<-1/(1+exp(-A[2]))

pp[1]<-qq[1]
pp[2]<-qq[2]-qq[1]
pp[3]<-1-qq[2]

# Proportion positive FISH result per IHC category
rr[1]<-1/(1+exp(-B[1]))
rr[2]<-1/(1+exp(-B[2]))
rr[3]<-1/(1+exp(-B[3]))

# Sensitivity per IHC category
S[1]<- 1
S[2]<- (pp[2]*rr[2]+pp[3]*rr[3])/(pp[1]*rr[1]+pp[2]*rr[2]+pp[3]*rr[3])
S[3]<- pp[3]*rr[3]/(pp[1]*rr[1]+pp[2]*rr[2]+pp[3]*rr[3])
S[4]<- 0

# Specificity per IHC category
C[1]<- 0
C[2]<- (1-rr[1])*pp[1]/((1-rr[1])*pp[1]+(1-rr[2])*pp[2]+(1-rr[3])*pp[3])
C[3]<- ((1-rr[1])*pp[1]+(1-rr[2])*pp[2])/((1-rr[1])*pp[1]+(1-rr[2])*pp[2]+(1-rr[3])*pp[3])
C[4]<- 1

}
```

Chapter 7

General discussion

A key step in any diagnostic accuracy study is to determine whether the target condition is present or absent in each patient. This process is also known as verification. Ideally, there is a gold standard that provides error-free classification. Accuracy measures, such as test sensitivity, specificity, likelihood ratios, predictive values, or diagnostic odds ratios, express how well the results of the test(s) under evaluation agree with the outcome of that gold standard.[1;3]

In most, if not all, cases a gold standard without error or uncertainty does not exist.[4;14;15] In these circumstances, researchers use the best available method to determine the presence or absence of the target condition, and refer to this as the clinical reference standard rather than gold standard.[3;16;17] Even within this framework, several problematic verification situations can occur. It may not be possible to perform the preferred reference test in all patients or that test may be substantially imperfect.[3] In applied clinical research, in spite of the presence of verification problems, naive analyses are still being performed and the verification problems are, at best, only briefly mentioned in the discussion section of the papers.

In the more technical literature, multiple solutions have been proposed to alleviate the bias caused by various reference standard problems.[2;5;9;10;18;19] These solutions are not frequently applied, as they are often rather technical and limited guidance is provided which solution should be preferred when, as all solutions have their advantages and disadvantages.

In this discussion, we will give an overview of the reference standard situations that may occur in studies of diagnostic accuracy. By using the ideal reference standard situation as a starting point, we classify possible reference test problems against two main axes: imperfection of the reference test and the completeness of the verification by this reference test. We then use this classification to provide guidance for researchers that have to deal with specific verification problems or for readers when reading reports about studies with such problems. In a flowchart we will summarize the verification problems and provide key questions for selecting possible solutions to alleviate these problems.

Ideal reference standard situation in diagnostic accuracy studies

Ideally, all subjects in a diagnostic accuracy study are verified by a single and (near) perfect reference standard. The various measures of accuracy of the index test(s) under study can then be calculated in a straightforward manner using classical methods.[1;20]

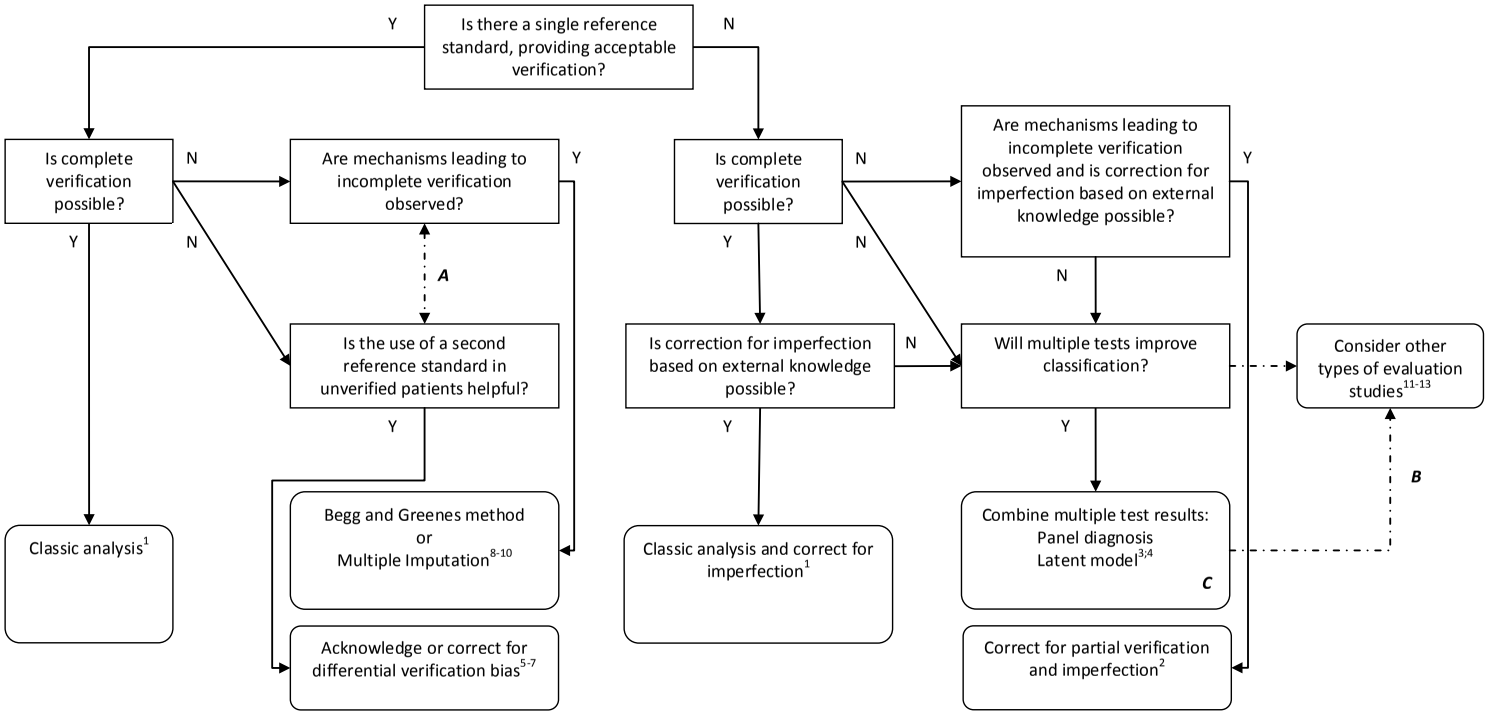

Figure 1. Updated guidance for researchers faced with problematic reference standard situations in diagnostic accuracy studies.

Two main axes of reference standard problems: incomplete verification and imperfection
Using this ideal situation as a starting point we can define two main categories of reference standard problems: incomplete verification and imperfection.
Based on these two axes, four main categories of reference standard situations can be defined:

1. ***Complete*** verification by a ***perfect*** reference standard
2. ***Incomplete*** verification by a ***perfect*** reference standard
3. ***Complete*** verification by an ***imperfect*** reference standard
4. ***Incomplete*** verification by an ***imperfect*** reference standard

The first reference standard situation is the ideal reference standard situation in diagnostic accuracy studies as mentioned earlier (Figure 1, vertical axis 1). We will now discuss the 3 problematic reference standard situations in more detail.

Incomplete verification by a perfect reference standard
If there is a single reference standard available that provides acceptable verification, but this reference standard can not be administered to all patients (Figure 1, vertical axis 2), we have a situation known as partial or incomplete disease verification. The bias associated with this is called partial verification bias, work-up bias, or referral bias.[1;6;21]

Different mechanisms can lead to partial verification.[3;6] Sometimes partial verification is simply unavoidable. For example, in a study using FDG-PET scanning to detect possible distant metastases before planning major curative surgery in patients with carcinoma of the oesophagus: only PET hot spots can be biopsied and verified histologically.[22] Incomplete verification can also be planned for efficiency reasons. This is often the case in screening test evaluation studies, where disease prevalence is low. In these types of studies, researchers often decide to apply the reference standard in only a random sample of the large group of patients with a negative screening test result. In other studies, partial verification is not planned, and reasons are unclear and not documented.[23;24]

There are two main strategies to address incomplete verification by the reference standard.

Mathematical methods to correct for incomplete verification
Mathematical methods to correct for partial verification bias, such as the Begg and Greenes method[8] or multiple imputation,[9,10] can be used to alleviate the partial verification bias. To optimally apply these techniques to address partial verification, researchers must collect detailed data about all relevant variables that may have been used in the (selective) referral of study subjects for verification. The performance of the correction methods will improve

with more and better information that may be involved in disease verification decisions. In short, the mechanisms leading to incomplete verification should be observed.

Introducing a second reference standard in unverified subjects

Another frequently encountered approach in diagnostic accuracy studies is to use an alternative, second best, reference test in those subjects where the result of the first, preferred reference test is absent. Although this seems a clinically appealing approach, bias can arise when the results of the two reference tests are treated as interchangeable. Both reference tests are almost by definition of different quality in terms of target disease classification or may even define the target disease differently.[25,26] Therefore, the results should be reported separately for each reference test to provide more clinically informative and unbiased measures of diagnostic accuracy.[7] If in these situations one still wants to quantify the accuracy of the diagnostic index test or model with regard to the same underlying target condition, one should also correct for possible imperfections of the applied reference tests.[5]

Complete verification by an imperfect reference standard

If there is no reference standard available that provides acceptable verification, but all subjects are verified by an imperfect reference standard (Figure 1, vertical axis 3), the question is whether correction for the imperfections of the reference standard is possible based on external knowledge. The fact that the imperfect reference standard is used as a reference test indicates that its properties with regard to the target condition could at least be approximately known from past experience, or by comparison with more accurate disease-detection methods (such as autopsy). If the degree of error of the used reference standard is known from previous studies or experience, the accuracy measures calculated by classic analyses can be mathematically adjusted to alleviate the bias due to the imperfection of the reference test.[16]

If correction for the imperfection of the reference standard is impossible or unfeasible, applying multiple diagnostic tests within patients can be used to improve classification. These tests often include general patient characteristics, signs and symptoms from history and physical examination, and other rest results.[4] These multiple tests need to be combined in order to classify subjects in those with and those without the target condition. Combining multiple test results can be done in several ways.

Panel diagnosis

In a panel diagnosis a group of clinical experts determines the presence or absence of the target condition in each patient based on the results of the available diagnostic tests. It is well known that there is an observed variation in physicians' practice-patterns for patients presenting with similar symptoms.[27] Within many medical specialties, doctors vary

substantially in observations, perceptions, and reasoning when deciding on appropriate medical care.[28] Recognizing the limitations of individual clinician diagnoses, research studies often use the consensus of a panel of multiple clinicians. It is hoped that a panel will achieve greater diagnostic reliability, accuracy, and certainty than an individual expert.[29]

Latent class analysis

Similar to clinicians combining several diagnostic test results to determine the presence or the absence of the target condition in a patient, one can also use a mathematical modelling technique to assess disease status. This is called a latent class analysis[30] and is based on the idea that observed variables are jointly determined by an underlying, unobserved (latent) target condition. In diagnostic accuracy studies the true disease status can be considered as a dichotomous latent variable with two categories, 'disease present' and 'disease absent'. Within a group of individuals with unknown disease status, for whom at least three independent diagnostic test results are available, both disease prevalence and sensitivity and specificity of all tests can be derived from the pattern of diagnostic tests results using latent class analysis.

Incomplete verification by an imperfect reference standard

The worst verification problems occur when in a diagnostic study only part of the research subjects are verified by an imperfect reference standard (Figure 1, vertical axis 4). If the mechanisms leading to incomplete verification and there is external knowledge on the imperfection of the reference standard available, one could use a mathematical correction method to correct for both the partial verification and the imperfection of the reference standard.[2]

If such information is unavailable, again multiple diagnostic tests could be used to improve classification. Like we previously described, the combination of the diagnostic tests as a method of verifying disease status can be done using either a consensus panel or latent class analysis. In more complex reference standard situations, as in studies with multiple imperfect or incomplete reference standards, the choice of an appropriate correction method or combination can be very difficult. In some situations, it might be preferable to acknowledge the fact that a standard diagnostic accuracy studies can or should not be performed and another type of evaluation studies, clinical test validation, should be considered.[3,4] We will address this other type of evaluation more thoroughly in the final section below.

Key remaining problems and future research

We have presented an overview of the possible reference standard situations that may occur in clinical practice. For some straightforward reference standard situations, such as in studies with a single (near) perfect but incomplete reference standard, mathematical correction methods to correct for the partial verification bias are available and fairly straightforward. In situations where verification by the reference standard is impossible or unethical in specific groups, researchers should beforehand think about the strategy for reference data collection and analysis. In our flowchart we specified the various forms and combinations of reference standard situations and possible methods to address and overcome the associated verification problems. However, several complex reference standard situations exist where it remains difficult to provide guidance which solution to prefer. We will briefly highlight a few of these situations.

The first difficult choice to make is whether to prefer a mathematical correction method or to introduce a second reference standard in unverified subjects when partial verification is present (Figure 1, ***A***). One important difficulty is that both methods are based on (different) assumptions that cannot be directly checked. In mathematical solutions the key assumption is that the mechanisms leading to partial perfection are observed. This "missing at random" assumption can never be formally demonstrated but the numerous simulation studies in this area have indicated that the quality of the imputation will improve by measuring all possible variables that might play a role in the decision of physicians to verify patients. Introducing an alternative, second best reference test in the unverified patients brings the problem of imperfection into the equation. In situations where partial verification is unavoidable and limited to a specific subgroup of the study population, reporting the results of the second reference standard for that subgroup alone (like a predictive value measure) is a better approach than combining the two reference standard results in order to produce overall measures like sensitivity, specificity or overall accuracy.

Because both methods have their merits and pitfalls, an empirical study directly comparing both approaches in the same set of patients seems valuable. Such a direct comparison will be informative when differences are small as researchers can then choose their preferred method. However, if the number of disagreements is substantial it will be impossible to determine which method is best without the help of an external, fair referee. This requires additional information. One source of additional information is to re-evaluate the disagreements to new clinical experts, but this solution will likely favor the panel approach. A more neutral approach is to use additional follow-up in the discordant patients to see whether complaints become clearer or whether specific events occur that increase the certainty that the condition of interest is present or absent. Another approach is to treat all discordant patients as if the condition is present and then examine whether one approach

is better than the other in predicting treatment response. More studies of this kind are needed to enable us to make more robust statements about which approach should be preferred in what clinical scenario.

Another problem is that in really complex reference standard situations it might be preferable to acknowledge the fact that standard diagnostic accuracy studies can or should not be performed and another type of evaluation studies, clinical test validation, should be considered (Figure 1, ***B***).[3,4] For example in studies where there is incomplete verification, correction methods can not be applied and second or multiple reference standards do not improve classification. No clear rules are defined about when the diagnostic accuracy paradigm can be preserved using correction methods or multiple reference standards or when researchers should choose for other types of evaluation studies. More clear definitions about when to leave the diagnostic accuracy paradigm and go for validation studies should be further investigated.

Finally, uncertainty also exist about the best method for combining different test results that have been collected within patients: a consensus panel diagnosis or a latent class model (Figure1, ***C***). A panel diagnosis seems to be a clinically appealing approach, but theoretical and empirical studies of group decision making indicate that, depending on their composition and procedures, consensus panels may not always achieve highly accurate decisions.[31] The efficacy of a panel varies with group size, heterogeneity of experts, and competence of the experts. A latent class analysis is a statistically sound and flexible approach and its potential benefit is that it is more objective than a panel diagnosis as it is examining the strength of statistical relationships among variables.[3] Yet, since disease is defined in statistical terms clinicians can feel uncomfortable what the results stand for. Furthermore, there can be technical problems when constructing a proper latent class models. It can be difficult to choose the correct way to model the dependence between variables, model convergence problems may occur and the statistical classification may not coincide with pre-existing knowledge of the target condition or it may even refer to a related, but different condition. In Chapter 5 we introduced the idea to do head-to-head comparisons between a panel diagnosis and latent class models. Again the difficult issue in such studies will be the choice of a fair referee in case of disagreements, as we discussed before. This idea should be further developed and applied in the near future.

In the diagnostic literature, naive analyses are still common despite various verification problems being present. We hope that this overview will urge researchers and readers to be more aware of reference standard problems and, if present, will use the most appropriate method to alleviate the bias. Our flowchart can be helpful in selecting the best available solution depending on the characteristics of the verification problem at hand.

Reference List

1. Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic procedures. The evidence base of clinical diagnosis. London: BMJ Books; 2002. 1-18.
2. 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(24):2532-2543.
3. 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(8):797-806.
4. 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):iii, ix-51.
5. de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Bossuyt PM, Moons KG. Adjusting for Differential-verification Bias in Diagnostic-accuracy Studies: A Bayesian Approach. Epidemiology 2011; 22(2):234-241.
6. de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N., Janssen KJ et al. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ. In press 2011.
7. Rutjes AW, Reitsma JB, Irwig LM, Bossuyt PM. Sources of bias and variation in diagnostic accuracy studies. In: AWS Rutjes, editor. Partial and differential verification in diagnostic accuracy studies. Amsterdam: Rutjes; 2005. 31-44.
8. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39(1):207-215.
9. de Groot JA, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB. Multiple imputation to correct for partial verification bias revisited. Stat Med 2008; 27(28):5880-5889.
10. de Groot JA, Janssen KJ, Zwinderman AH, Bossuyt PM, Reitsma JB, Moons KG. Correcting for partial verification bias: a comparison of methods. Ann Epidemiol 2011; 21(2):139-148.
11. Streiner DL, Norman GR. “Precision” and “accuracy”: two terms that are neither. J Clin Epidemiol 2006; 59(4):327-330.
12. Gagnon R, Charlin B, Coletti M, Sauve E, Van d, V. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? Med Educ 2005; 39(3):284-291.
13. Bland JM, Altman DG. Statistics Notes: Validating scales and indexes. BMJ 2002; 324(7337):606-607.
14. Valenstein PN. Evaluating diagnostic tests with imperfect standards. Am J Clin Pathol 1990; 93(2):252-258.

15. Pepe MS. Incomplete data and imperfect reference tests. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2004. 168-213.
16. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, editor. The Evidence Base of Clinical Diagnosis. 2nd ed. London: BMJ Books; 2002. 39-60.
17. 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. Ann Intern Med 2003; 138(1):W1-12.
18. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. Biometrics 2003; 59(1):163-171.
19. Kosinski AS, Barnhart HX. A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. Stat Med 2003; 22(17):2711-2721.
20. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ 1994; 308(6943):1552.
21. Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987; 6(4):411-423.
22. Lee J, Aronchick JM, Alavi A. Accuracy of F-18 fluorodeoxyglucose positron emission tomography for the evaluation of malignancy in patients presenting with new lung abnormalities: a retrospective review. Chest 2001; 120(6):1791-1797.
23. Pode D, Shapiro A, Lebensart P, Meretyk S, Katz G, Barak V. Screening for prostate cancer. Isr J Med Sci 1995; 31(2-3):125-128.
24. Elhendy A, van Domburg RT, Poldermans D, Bax JJ, Nierop PR, Geleijnse ML et al. Safety and feasibility of dobutamine-atropine stress echocardiography for the diagnosis of coronary artery disease in diabetic patients unable to perform an exercise stress test. Diabetes Care 1998; 21(11):1797-1802.
25. 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(3):189-202.
26. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11):1061-1066.
27. Elmore JG, Feinstein AR. A bibliography of publications on observer variability (final installment). J Clin Epidemiol 1992; 45(6):567-580.
28. Eddy D. Clinical Decision Making. London: Jones and Bartlett Publishers; 1996.
29. Gabel MJ, Foster NL, Heidebrink JL, Higdon R, Aizenstein HJ, Arnold SE et al. Validation of consensus panel diagnosis in dementia. Arch Neurol 2010; 67(12):1506-1512.

30. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. Biostatistics 2007; 8(2):474-484.
31. Gabel MJ, Shipan CR. A social choice approach to expert consensus panels. J Health Econ 2004; 23(3):543-564.

Summary

Diagnostic accuracy research is a vital step in the evaluation of new diagnostic technologies. It is the ability of a test to correctly discriminate between patients that have and do not have the target disease. In studies of diagnostic accuracy, results of the tests under study are compared with results of a reference standard applied to the same patients. In this research framework, the reference standard is the best available method to verify the presence or absence of the target disease, and thus provides the final classification of patients into target disease present or absent. This process is known as disease verification.

Ideally, the reference standard provides error-free disease classification. In some situations, it is not possible to verify the disease outcome with the preferred reference standard in all patients or any patient at all. Failure to apply the reference standard may result in various types of disease verification problems. Biased and exaggerated estimates of accuracy of a test can lead to inefficiencies in testing in clinical practice, unnecessary costs, and could trigger physicians to making incorrect treatment decisions.

This thesis examines the problem of verification bias in studies of diagnostic accuracy. In particular, we aimed to investigate the available methods to alleviate the various problems of verification bias and, more importantly, to improve the methodology and analysis of primary diagnostic accuracy studies and diagnostic meta-analyses in the presence of various forms of verification bias.

In **Chapter 2** we describe two important types of disease verification problems, partial – and differential verification, using empirical examples and propose solutions to alleviate the associated biases. Partial verification bias occurs when not all patients are verified by the reference standard, and when the referral for disease verification is related to other factors like other test results or patient characteristics. If the preferred reference standard has not been applied in all patients, selectively or non-selectively, mathematical correction methods can be used to correct for the partial verification bias. Another approach in diagnostic accuracy studies is to use an alternative reference test in those subjects where the result of the first, preferred reference test cannot be obtained. Although this seems a clinically appealing and ethical approach, differential verification bias arises when the results of both reference tests are treated as equal and interchangeable when, in fact, they are of different quality or define the target condition differently. Then, the estimated accuracy of the diagnostic index test or model should be corrected and reported separately for each reference test, to provide more informative and less biased index tests' accuracy measures.

Chapter 3 focuses on the problem of partial verification bias. In **Chapter 3.1** we revisit a previous publication on multiple imputation to correct for partial verification bias. In that article the authors showed that partial verification can be considered as a missing data problem and that multiple imputation methods can be used to correct for this bias. They claim that even in simple situations where the verification is random within strata of the index test results, the so-called Begg and Greenes correction method underestimates sensitivity and overestimates specificity as compared with the MI method. However, we were able to demonstrate that the B&G method produces similar results as MI, and that the claimed difference has been caused by a computational error. In **Chapter 3.2** we compare the performance of multiple imputation and the conventional correction method proposed by Begg and Greenes ourselves, under a range of different situations of partial verification. In a series of simulations, using a previously published Deep Venous Thrombosis dataset (N=1292), we set the outcome of the reference standard to missing based on various underlying mechanisms and by varying the total number of missing values. We then compared the performance of the different correction methods. The results of this study show that when the mechanism of missing reference data is known, accuracy measures can easily be correctly adjusted using either the Begg and Greenes method, or multiple imputation. In situations where the mechanism of missing reference data is complex or unknown, we strongly recommend using multiple imputation methods to correct. Multiple imputation methods can easily be extended with more variables, can incorporate both continuous and categorical variables, and they are readily available in statistical software.

Chapter 4 focuses on the problem of differential verification bias. In **Chapter 4.1** we present Bayesian methods using a single model to 1) acknowledge the different nature of the two reference standards applied, 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.

In **Chapter 5** we introduce the idea to directly compare two different verification strategies to handle diagnostic accuracy studies when there is no generally accepted reference standard available, and researchers have collected multiple test results on each patient: the consensus panel and the latent class model. We illustrate the application of both methods for combining various test results using the data from a heart failure study and discuss the possible origin of observed differences in results. Although estimates of accuracy of individual tests were highly comparable, on a patient level there was a fair amount of

disagreement (20.2% of all patients). A universal problem in such comparison studies is how to decide which method is correct if the two approaches disagree in their classification. Ideally, a third external method exists which can act as a fair referee. More extensive follow-up with additional testing could be an option. The results of this case study may generate further ideas for analysing and designing studies to critically compare different solutions for diagnostic accuracy studies with verification problems.

Numerous methods have been described for correcting partial verification bias in individual studies. In **Chapter 6** we describe a Bayesian method to obtain unbiased results from a diagnostic meta-analysis when partial verification or work-up bias is present in a subset of the primary studies. The method corrects for this 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 our method are compared to the existing methods for dealing with verification bias in diagnostic meta-analyses. Empirical data from a systematic review of studies of the accuracy of the immunohistochemistry test for diagnosis of HER-2 status in breast cancer patients is used for illustration.

Chapter 7 gives a final overview of the reference standard situations that may occur in studies of diagnostic accuracy. By using the ideal reference standard situation as a starting point, we classify possible reference test problems against two main axes: imperfection of the reference test and the completeness of the verification by this reference test. We then use this classification to provide guidance for researchers that have to deal with specific verification problems or for readers when reading reports about studies with such problems. In a flowchart we summarize the verification problems and provide key questions for selecting possible solutions to alleviate these problems. We hope that this overview will urge researchers to think about their reference standard situations and, if possible, use available methods to alleviate the corresponding bias.

Nederlandse samenvatting

Diagnostisch accuratesse onderzoek is een essentiële stap in de evaluatie van nieuwe diagnostische technologieën. Diagnostische accuratesse geeft aan hoe goed een test in staat is om onderscheid te maken tussen patiënten met en zonder een bepaalde ziekte. Om deze diagnostische accuratesse te bepalen, worden de resultaten van een nieuwe diagnostische test vergeleken met de resultaten van een referentiestandaard binnen eenzelfde groep van patiënten. De referentiestandaard is de best beschikbare test om de aan- of afwezigheid van een specifieke ziekte aan te tonen. Dit aantonen of uitsluiten van de ziekte via de referentiestandaard heet ook wel verificatie.

Idealiter kan de referentiestandaard foutloos onderscheid maken tussen patiënten met en zonder een bepaalde ziekte of aandoening. Soms is het echter niet mogelijk om alle patiënten in een onderzoek de referentiestandaard te laten ondergaan, bijvoorbeeld wanneer de gewenste referentiestandaard te duur of te invasief is.

Als het niet mogelijk is om alle patiënten in een onderzoek de referentiestandaard te laten ondergaan, leidt dit vaak tot bias in de schattingen van de accuratesse. Dit verkeerde beeld over een test kan diverse negatieve gevolgen hebben, zoals onnodige kosten door inefficiënt gebruik van de test, tot zelfs incorrecte beslissingen over de behandeling.

In dit proefschrift beschrijven we diverse methoden om de gevolgen van problemen in de verificatie te verlichten. Deze methoden kunnen worden ingezet in zowel de analyse van primaire diagnostische accuratesse studies als binnen meta-analyses van diagnostische studies met verificatieproblemen.

In **hoofdstuk 2** beschrijven we aan de hand van empirische voorbeelden twee belangrijke vormen van verificatieproblemen; partiële- en differentiële verificatie. Partiële verificatie bias treedt op als niet alle patiënten in een onderzoek de referentiestandaard ondergaan en als de reden daarvan afhangt van andere patiëntkenmerken of eerdere testresultaten. Er bestaan mathematische correctiemethoden om dit probleem te verlichten. Een andere aanpak voor dit verificatieprobleem is echter om bij patiënten die de gewenste referentiestandaard niet hebben ondergaan, een tweede alternatieve referentietest uit te voeren. Dit lijkt een klinisch logische benadering, maar er kan bias ontstaan als de twee verschillende referentietesten als gelijk en uitwisselbaar worden beschouwd. De bias die hierdoor kan ontstaan, heet differentiële verificatie bias. Indien twee (of meer) verschillende referentietesten worden gebruikt, dient de geobserveerde accuratesse van een nieuwe diagnostische test modelmatig te worden gecorrigeerd en per gebruikte referentietest te worden gerapporteerd.

Hoofdstuk 3 focust zich op de bias door partiële verificatie. In **hoofdstuk 3.1** bespreken we een eerder gepubliceerd artikel over multiple imputatie als correctiemethode voor partiële verificatie bias. In dat artikel laten de auteurs zien dat partiële verificatie kan worden gezien als een 'missing data' probleem en dat multiple imputatie kan worden gebruikt ter correctie voor deze vorm van bias. De auteurs betogen dat zelfs in eenvoudige situaties waar verificatie volledig random is binnen strata van resultaten van de indextest, multiple imputatie tot betere resultaten leidt dan een andere bekende correctiemethode beschreven door Begg and Greenes. Echter, wij tonen in dit hoofdstuk aan dat de Begg and Greenes methode tot vergelijkbare resultaten leidt als multiple imputatie en dat het eerder beschreven verschil voortkwam uit rekenfouten. In **hoofdstuk 3.2** vergelijken we zelf de prestaties van multiple imputatie en de conventionele correctiemethode van Begg en Greenes in verschillende scenario's van partiële verificatie. In een serie van simulaties in een complete dataset over diep veneuze trombose (N=1292) hebben we selectief uitkomsten van referentietest missing gemaakt op basis van verschillende onderliggende mechanismen en variërend in aantal. Daarna hebben we de verschillende correctiemethoden met elkaar vergeleken. De resultaten laten zien dat als het mechanisme van het ontstaan van missende data bekend is (missing at random), zowel multiple imputatie als de Begg en Greenes methode prima als correctiemethode gebruikt kunnen worden. Beide methoden presteren in dat geval even goed. In situaties waar het mechanisme van het ontstaan van missende referentiestandaard resultaten onbekend of erg complex is, adviseren we het gebruik van multiple imputatie. Multiple imputatiemethoden hebben namelijk als voordeel dat zij eenvoudig uitgebreid kunnen worden met meer variabelen, die zowel continu als categorisch kunnen zijn. Bovendien zijn multiple imputatiemethoden tegenwoordig beschikbaar in diverse statistische softwarepakketten.

In **hoofdstuk 4** staat het probleem van differentiële verificatie bias centraal. In **hoofdstuk 4.1** presenteren we een Bayesiaans model dat tegelijkertijd 1) de verschillen van de twee gebruikte referentietesten in acht neemt, en 2) schattingen genereert van de prevalentie, sensitiviteit en specificiteit zowel afgezet tegen beide gebruikte referentietesten afzonderlijk als tegen een statistisch geconstrueerde indeling van zieken en niet-zieken (latente klasse model). We illustreren deze methode door de data van een studie naar de accuratesse van het strekken van de elleboog voor het diagnosticeren van een elleboogfractuur te heranalyseren. In deze studie vond verificatie plaats door röntgenfoto's dan wel door het beloop van klachten in de follow-up.

In **hoofdstuk 5** vergelijken we twee verschillende strategieën om in situaties waar geen geaccepteerde referentiestandaard bestaat toch de diagnostische accuratesse van een nieuwe test te onderzoeken: een consensus uitspraak door een panel van deskundigen

versus een latente klasse model. We passen beide methoden toe binnen een diagnostische studie naar hartfalen en bediscussiëren de mogelijke herkomst van de verschillen tussen beide methoden. Hoewel de berekende accuratesse van de individuele testen sterk overeenkwamen, waren er op het niveau van de patiënt aanzienlijke verschillen (tot in 20.2% van de patiënten). Een universeel probleem in dit soort vergelijkende studies is hoe te beslissen welke methode correct is als de resultaten verschillen. Idealiter bestaat er een derde, onafhankelijke methode van hoge kwaliteit die als scheidsrechter kan optreden. Zorgvuldige follow-up waarin het beloop van klachten wordt vastgelegd, eventueel aangevuld met extra testen, zou een goede methode hiervoor kunnen zijn.

Er bestaan verschillende methoden om voor partiële verificatie in individuele studies te corrigeren. In **hoofdstuk 6** beschrijven we een Bayesiaanse methode om binnen een diagnostische meta-analyse te corrigeren voor partiële verificatie die aanwezig is in een deel van de geïncludeerde primaire studies. Deze methode corrigeert voor de bias, zonder de primaire studies met partiële verificatie volledig uit de analyse te schrappen. Hierdoor blijven twee belangrijke eigenschappen van een meta-analyse overeind: verbeterde precisie en generaliseerbaarheid. De resultaten van deze aanpak worden vergeleken met bestaande correctiemethoden voor partiële verificatie in diagnostische meta-analyses. We gebruiken de data van een systematische review naar de accuratesse van een immunohistochemie test voor het diagnosticeren van de HER-2 status in patiënten met borstkanker om onze methode te illustreren.

Hoofdstuk 7 geeft een overzicht van de diverse verificatieproblemen die kunnen voorkomen in diagnostisch accuratesse onderzoek. Met het ideale diagnostische design als startpunt, classificeren we deze problemen aan de hand van twee assen: imperfectie van de referentietest en de compleetheid van verificatie met de referentietest. Zowel onderzoekers als lezers van diagnostische artikelen kunnen deze assen als hulpmiddel gebruiken om beter zicht te krijgen op eventueel aanwezige verificatieproblemen. We vatten de verschillende problemen rondom de verificatie en de mogelijke manieren om de bias te verlichten in een stroomdiagram samen. We hopen dat dit overzicht onderzoekers aanspoort om scherper na te denken over de verificatie binnen hun studie en, wanneer mogelijk, beschikbare correctiemethoden te gebruiken.

Dankwoord

Hoewel promoveren te boek staat als een erg solistisch traject, hebben vele mensen meegeholpen aan de totstandkoming van deze thesis. Ik wil graag van de gelegenheid gebruik maken om hen te bedanken voor hun steun, inzet en vertrouwen.

Prof. dr. K.G.M. Moons, beste Carl. Ik had me geen betere promotor kunnen wensen. Je bent altijd enthousiast over alle projecten, waardoor je me steeds weer motiveert om iets nieuws uit te proberen, dingen uit te zoeken of verder te schrijven aan onze artikelen. Ik heb er grote bewondering voor dat je altijd bereikbaar bent geweest, ook op momenten als het even tegenzat. Ik kijk ernaar uit om samen de komende jaren de methoden van diagnostisch onderzoek nog verder te ontwikkelen en verbeteren.

Prof. dr. P.M.M. Bossuyt, beste Patrick. Ik heb veel respect voor jouw ervaring en inzicht op het gebied van de klinische epidemiologie. Je snelle, scherpe en steekhoudende commentaren en aanvullingen op mijn papers, waren essentieel voor de kwaliteit daarvan. Bedankt voor je begeleiding en kritische inbreng bij het schrijven van dit proefschrift.

Dr. K.J.M. Janssen, beste Kristel. Hoewel je twee keer met zwangerschapsverlof was en uiteindelijk een nieuwe uitdaging hebt gevonden, was je een begenadigd dagelijks begeleidster. Erg fijn hoe laagdrempelig het was om bij je binnen te lopen met mijn vragen en problemen, zodat ik weer verder kon. Ik zal voor altijd zowel het outputvenster als het scriptvenster onder elkaar in R gebruiken. Heel veel plezier en succes met je nieuwe uitdaging en natuurlijk je jonge gezinnetje.

Dr. J.B. Reitsma, beste Hans. Je bent hét voorbeeld van een clinicus met een passie voor methodologie. Deze combinatie maakt dat je vaak lastige theoretische, methodologische vraagstellingen moeiteloos kan vertalen naar de relevante klinische praktijk. Het is zeer prettig om met je samen te werken en ik ben dan ook blij dat je ons team in het UMC Utrecht bent komen verrijken. Ik hoop nog vele jaren van je te kunnen leren.

Dr. N. Dendukuri, dear Nandini. Thank you for giving me the opportunity to visit you in Montreal. I am very grateful to you, not only because the 6 months I spent at McGill were a great intellectual experience and a great opportunity to learn from your extensive knowledge about Bayesian analyses, but even more so because you took the time to personally guide me on a daily basis. Nandini, you are a wonderful person and you made my stay in cold Montreal a warm and great experience. Thank you so much and I am happy that our cooperation will continue in the future.

Dr. Madhukar Pai, thank you for helping me with the preparations for my research stay in Montreal. Prof. dr. Lawrence Joseph and dr. James Brophy, it was very inspiring to meet you. Thank you for co-authoring our meta-analysis paper.

De beoordelingscommissie bestaande uit Prof. dr. M.L. Bots, Prof. dr. S. van Buuren, Prof. dr. C.J. Kalkman en Prof. dr. M.H. Prins dank ik voor hun bereidheid dit manuscript te lezen en te beoordelen.

Uiteindelijk heb ik meer dan 3 jaar achter verschillende bureaus met verschillende kamergenoten op kamer 6.119 gezeten. Eerst met Auke, Hadi, Annemieke, Rieke, Saskia, Carianne, Arnoud en Peter en later met Sjoukje, Tannie, Lisette, Liselotte, Nanne, Gerdien, Paulien, Thomas, Stan en Floriaan. Dank voor al jullie gezelligheid, collegialiteit en inspirerende gesprekken. En natuurlijk voor het bewaken van mijn bureau als ik er weer eens niet was.

Annina en Coby, bedankt voor jullie hulp bij alle soms lastige praktische klusjes. Mijn proefschrift mag dan een blijk van theoretisch vermogen zijn, organisatorisch schiet ik nogal eens te kort. Ik hoop ook in de toekomst nog vaak van jullie kwaliteiten te mogen profiteren.

Omdat namen als Borstel, Dinho, Djek, Longley, Carlos, King en Rat mijn proefschrift wel heel erg zouden vervuilen, schaar ik dezen en vele anderen maar onder de gezamenlijke noemer vrienden. Ik bedank jullie voor het feit dat jullie me er constant aan herinneren dat er meer is in het leven dan onderzoek doen. Met jullie heb ik het nagenoeg nooit over mijn werk, wat heerlijk is dat! Met zijn allen of in kleinere teams een balletje gooien, een kaartje leggen, een klein drankje doen, naar een fiske, op reis etc. etc. Al dat soort dingen hebben mij de benodigde energie gegeven om mijn serieuzere werkzaamheden te kunnen doen.

Ron, ik neem aan dat ook jij nog steeds niet goed begrijpt waar ik eigenlijk de voorbije vier jaar mee bezig ben geweest. En toch heb je alles nauwlettend op de voet gevolgd en steeds geïnteresseerd geluisterd naar wat ik te vertellen had. Hoewel het fenomeen paranimf voor jou nog onbekend was, wist ik al vanaf het begin dat deze rol tijdens mijn promotie jou toekomt. Ik vind het een eer dat je ook bij deze gelegenheid achter me staat.

Verder wil ik natuurlijk mijn familie en schoonfamilie danken voor hun interesse en steun tijdens mijn promotietraject. Ik vind het erg bijzonder dat ik jaren na dato in dezelfde senaatzaal sta als waar mijn lieve oma mijn ‘Oom Antoon’ al zijn bul in ontvangst zag nemen. Fijn om dit stukje familiehistorie voort te kunnen zetten.

Lieve papa en mama, van alle personen ben ik jullie wel het allermeeste dank verschuldigd. Het lijkt misschien soms alsof ik het als vanzelfsprekend beschouw wat jullie allemaal voor mij gedaan hebben en nog steeds doen, maar dat doe ik zeker niet. Jullie hebben mij de mogelijkheid en vrijheid gegeven om mij te ontwikkelen tot die persoon die ik nu ben. Ik heb gedurende mijn studies en promotieonderzoek altijd op jullie onvoorwaardelijke, positieve en stimulerende steun kunnen rekenen. Ik heb het gevoel dat jullie trots zijn op wat ik bereikt heb. Ben vooral erg trots op jullie zelf, want zonder jullie had ik dit nooit bereikt. Ik ben erg trots dat papa namens jullie beiden mijn paranimf wil zijn.

Erwin en Sandra, ik ben zo blij dat ik er met jullie als mijn “grote” broer en zus nooit alleen voorsta. Hoewel we elkaar soms te weinig zien of spreken, weet ik dat ik altijd op jullie kan rekenen. Ik ben trots op jullie en ik vind dat we vaker met elkaar een hapje moeten gaan eten! En Jasper en Jikke, wat fijn dat jullie meer en meer een extra broer en zus voor me worden.

Als allerlaatste bedank ik mijn schattie. Lieve Marjon, ik ben zo blij dat we elkaar tijdens mijn promotietraject eindelijk hebben gevonden. Hoewel ik vreesde dat het combineren van ons eigen huissie, inclusief flinke verbouwing, met het afronden van mijn promotietraject niet bepaald ideaal zou zijn, bleek het samen met jou een fluitje van een cent. Ik ben zó blij met jou, jij maakt me gelukkig!

Curriculum Vitae

Joris Antoon Harry de Groot was born on July 30, 1980 in Rosmalen, the Netherlands.

As of 1998 he studied Biomedical Health Sciences and medicine at the Radboud University Nijmegen. From February to August 2002 he worked as an intern researcher at The Miami Project to Cure Paralysis, Miller School of Medicine, University of Miami, Miami (FL, USA) focusing on postprandial lipemia in chronic paraplegia. In 2005 he graduated as Master of Science in Biomedical Health Sciences with major subjects exercise physiology and neurology.

From 2007 till 2011 Joris de Groot performed his PhD studies as a researcher at both the Julius Center for health sciences and primary care, University Medical Center Utrecht (the Netherlands) and at the Academic Medical Center Amsterdam (the Netherlands), which has resulted in the studies presented in this thesis. He was supervised by Prof. dr. K.G.M. Moons, Prof. dr. P.M.M. Bossuyt, Dr. J.B. Reitsma and Dr. K.J.M. Janssen. During his PhD studies, Joris de Groot obtained his MSc in Clinical Epidemiology at Utrecht University based on his research on Bayesian methods to correct for differential verification bias. Notably, with this research Joris de Groot won the Professor Frits De Waard Award 2011 for the best and most original epidemiologic research by a student in Biomedical Sciences. Moreover, during January-August 2009 he worked as a researcher at the Department of Epidemiology and Biostatistics at McGill University, Montreal (CA), under supervision of Dr. Nandini Dendukuri.

Currently, Joris de Groot works as post-doctoral researcher on methodological aspects of diagnostic research at the Julius Center of health sciences and primary care, University Medical Center Utrecht, the Netherlands.