Journal of Clinical Epidemiology 111 (2019) 1−10

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Forcing dichotomous disease classification from reference standards leads to bias in diagnostic accuracy estimates: A simulation study

Kevin Jenniskens[a,*], Christiana A. Naaktgeboren[a], Johannes B. Reitsma[a,b], Lotty Hooft[a,b], Karel G.M. Moons[a,b], Maarten van Smeden[a,c]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[b]*Cochrane Netherlands, University Medical Center Utrecht, Utrecht, The Netherlands*
[c]*Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands*

Accepted 1 March 2019; Published online 20 March 2019

Abstract

Objectives: The objective of this study was to study the impact of ignoring uncertainty by forcing dichotomous classification (presence or absence) of the target disease on estimates of diagnostic accuracy of an index test.

Study Design and Setting: We evaluated the bias in estimated index test accuracy when forcing an expert panel to make a dichotomous target disease classification for each individual. Data for various scenarios with expert panels were simulated by varying the number and accuracy of ''component reference tests'' available to the expert panel, index test sensitivity and specificity, and target disease prevalence.

Results: Index test accuracy estimates are likely to be biased when there is uncertainty surrounding the presence or absence of the target disease. Direction and amount of bias depend on the number and accuracy of component reference tests, target disease prevalence, and the true values of index test sensitivity and specificity.

Conclusion: In this simulation, forcing expert panels to make a dichotomous decision on target disease classification in the presence of uncertainty leads to biased estimates of index test accuracy. Empirical studies are needed to demonstrate whether this bias can be reduced by assigning a probability of target disease presence for each individual, or using advanced statistical methods to account for uncertainty in target disease classification.

Keywords: Imperfect reference standard; Bias; Expert panel; Diagnostic test accuracy studies; Dichotomization; Simulation study

1. Introduction

In diagnostic test accuracy studies, the discriminatory ability of the test of interest (index test) is evaluated by comparing its results with those of the reference standard in a group of individuals suspected of the target disease. While analyzing this comparison, it is often assumed that the reference standard can perfectly distinguish two groups of individuals: those with and without the target disease [1,2]. For many diseases, however, the best available reference standard is not perfect [3,4]. In the absence of a single perfect test that can be held as the reference standard, alternative approaches have been proposed, including composite reference standards (applying multiple tests and combining their results using a fixed rule), latent class models (using a statistical method to link multiple tests' results to a latent class), and expert panels [5].

Using an expert panel is a common approach to assign a final diagnosis in fields where an accepted reference standard is lacking [6]. In such a panel, multiple experts combine information from multiple tests, patient characteristics, and clinical expertise to make a final decision on whether the target disease is present or absent for each individual. Typically, in an expert panel, all individuals are ultimately classified as either having or not having the target disease based on a decision-making procedure, such as majority vote or by consensus [6,7]. With this dichotomization (presence or absence) of the target disease, measures of index test accuracy can be calculated in the traditional way [8,9].

Compared with a single-test error-prone reference standard, the panel diagnosis may improve reference standard

Funding Statement: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, PO Box 85500, 3508GA Utrecht, The Netherlands.

E-mail address: K.Jenniskens@umcutrecht.nl (K. Jenniskens).

https://doi.org/10.1016/j.jclinepi.2019.03.002
0895-4356/

What is new?

Key findings

- Ignoring uncertainty in target disease classification by a reference standard leads to biased estimates of sensitivity and specificity of the index test under study.

What this adds to what was known?

- A step-by-step illustrative example is used to guide readers through understanding the process that leads to bias in estimates of sensitivity and specificity when ignoring uncertainty in the target disease classification, a situation that is common in studies using an expert panel as reference standard.
- This adds to evidence of similar issues that can occur when a composite reference standard is used.

What is the implication and what should change now?

- Researchers involved in diagnostic studies using an expert panel or composite reference standard should consider obtaining estimates of target disease probability, or making use of advanced statistical methods such as latent class analysis, to account for diagnostic uncertainty.

accuracy and subsequently reduce reference standard bias [5,10]. However, panel diagnoses almost by definition lead to imperfect target disease classification, as evidenced by studies of panel intraobserver and interobserver variability [11−13]. Different experts within a panel can disagree on the presence/absence of the target disease, in particular in patients presenting with atypical signs and symptoms. Forcing a dichotomous decision in every individual thus ignores this uncertainty about the target disease status. Simply ignoring this uncertainty may lead to biased accuracy estimates of the index test under study. This has already been demonstrated for composite reference standards using explicit decision rules (e.g., at least two of four tests should be positive to assign a target disease present classification to an individual) [14,15], but has not been described in the context of expert panels, which is the goal of this article.

In this study, we aim to assess the impact of dichotomization of the target disease classification on accuracy estimates of the index test. An expert panel with multiple imperfect tests at its disposal will be used as a reference standard. We first present an example to illustrate how ignoring uncertainty in the target disease classification can lead to biased accuracy estimates of the index test. Readers familiar with this type of bias can skip this section (Section 2) and directly go to the description of the methods and results of our simulation study, illustrating the bias due to dichotomization of target disease status across a range of scenarios. Implications of the results for diagnostic research will be discussed, and alternative strategies for reducing bias in index test accuracy estimates will be proposed.

2. The source of bias: an illustrative example

Consider the following hypothetical example of 1,000 individuals with a target disease prevalence of 40% to which an index test with sensitivity and specificity 80% is applied. Assuming we have a perfect (gold) reference standard, we can construct Table 1, which we will refer to as the true contingency table.

Now suppose that there is no perfect reference standard, and that the disease classification is made by a panel of experts. These expert panels are frequently applied in various clinical domains such as psychiatric disorders and cardiovascular diseases when a single reference standard is lacking [6]. For example, when assessing screening tools for diagnosis of autism spectrum disorder, expert panels are used as a reference standard, which requires (subjective) interpretation of different components from the Diagnostic and Statistical Manual of Mental Disorders [16,17]. Note that in this article, we will not consider a continuous spectrum of target disease severity, but rather focus on expert panel's uncertainty regarding the presence or absence of a single well-defined target condition.

Expert panels combine the results of several imperfect tests to make the final classification whether the target disease is present or absent. Each separate test available to the expert panel will hereinafter be referred to as a "component test". We use the term component test in a broad sense, as any piece of information (e.g., patient characteristic, biomarker, imaging) that might help in making the disease classification. In this example, we use two dichotomous component tests, the first having a true sensitivity and specificity of 80%, and the second a sensitivity and specificity of 90%. For simplicity, we assume that the errors of these component tests are uncorrelated; in other words, the test results are conditional independent given the true target disease status [18].

We simulate the implicit decisions by an expert panel on target disease classification by making them explicit, solely based on the assigned probability of target disease presence given the component test results for any given individual. Individuals are then classified to target disease present (i.e., probability of disease presence of 50% or higher) or target disease absent (i.e., probability of disease presence below 50%). We assume that the panel is well calibrated (they assign correct target disease presence probabilities) and consistent (they apply the same threshold value of 50% across all individuals when dichotomizing). Ultimately, the panel is forced to classify each individual as either being disease present or disease absent. This final classification is used to calculate sensitivity and specificity of the index test.

Table 1. True contingency table for a hypothetical index test when compared to a gold standard

	Disease present (according to gold standard)	Disease absent (according to gold standard)	Total
Index test +	320	120	440
Index test −	80	480	560
Total	400	600	1,000

Because this is a simulation study, we know the true values of sensitivity and specificity of the index test, and therefore, the corresponding bias can be calculated.

In this example, there are four possible component reference test patterns (++, +−, −+, −−). The probability of observing a specific test pattern is given by the sensitivity (Se) and specificity (Sp) of the component tests, and the target disease prevalence (prev). When a ++ pattern is observed, the first and second component test are positive. This can occur in two ways: an individual has the disease and these are two true positive component test results [with probability: prev*Se_{comp1}*Se_{comp2}] or an individual does not have the disease and these are two false positive component test results [with probability: (1−prev)*(1−Sp_{comp1})*(1−Sp_{comp2})]. The total probability of observing pattern ++ is the sum of these two probabilities. This can be generalized for each possible component test pattern, obtaining the following formulas for the probability of each pattern:

For our illustrative example, in Table 2, we can see that the probability of observing a component test pattern with two positive test results is 30%, and within that component test pattern, there is 96% (not 100%) probability of truly having the target disease. Hence in a sample of 1,000 individuals, the expert panel will assign the target disease to all 300 individuals having the ++ pattern, of which only 288 would truly have the target disease.

In practice, the expert panel makes a dichotomous decision about the presence of the target disease for each individual based on the results of the two component reference tests. To reach this decision, the expert panel applies a threshold (either implicitly or explicitly) on the probability of target disease being present. An intuitive threshold for dichotomizing disease status would be 50%, such that each individual is classified to the most likely disease status (present or absent). In our example, this would result in all individuals with component test pattern ++ and −+ being classified as disease present, as their probabilities of having the disease are higher than 50% (96% and 60%, respectively). The remaining component test patterns have probabilities below the 50% threshold; consequently, individuals with these patterns will be classified as disease absent. We will refer to component test patterns above the threshold as dichotomous classification 1 (DC1) and under the threshold as dichotomous classification 0 (DC0).

In our illustrative example, the expected distribution of the 1,000 individuals can be calculated using the probability of observing a component test pattern and the probability of target disease presence (Table 2). In DC1 there are two test patterns (++ and −+), in which 420 (300 + 120) individuals are classified as target disease present, and of which 360 (288 + 72) are truly diseased. In the test patterns in DC0 (+− and −−), zero individuals are classified as diseased, whereas in reality, 40 (32 + 8) are truly diseased. From this, we can also derive that the prevalence according to the expert panel's classification has

Pattern (k)		*Formulas for the probability for observing each possible component test pattern*		
		Probability for diseased cases		*Probability for nondiseased cases*
1	++	prev*Se_{comp1}*Se_{comp2}	+	(1−prev)*(1−Sp_{comp1})*(1−Sp_{comp2})
2	−+	prev*(1−Se_{comp1})*Se_{comp2}	+	(1−prev)*Sp_{comp1}*(1−Sp_{comp2})
3	+−	prev*Se_{comp1}*(1−Se_{comp2})	+	(1−prev)*(1−Sp_{comp1})*Sp_{comp2}
4	−−	prev*(1−Se_{comp1})*(1−Se_{comp2})	+	(1−prev)*Sp_{comp1}*Sp_{comp2}

The probability of target disease presence given a component test pattern can be derived by using Bayes' theorem [19,20]. For pattern ++, this post-test probability is given by the probability of observing that pattern among diseased [probability: prev*Se_{comp1}*Se_{comp2}] divided by the total probability of getting that pattern [probability: prev*Se_{comp1}*Se_{comp2} + (1−prev)*(1−Sp_{comp1})*(1−Sp_{comp2})]. Applying this line of reasoning to all component test patterns leads to the following formulas:

Pattern (k)		*Formulas for the probability of disease presence within a component test pattern*		
		Probability for diseased cases		*Probability for observing component test pattern*
1	++	prev*Se_{comp1}*Se_{comp2}	/	(prev*Se_{comp1}*Se_{comp2}+(1−prev)*(1−Sp_{comp1})*(1−Sp_{comp2}))
2	−+	prev*(1−Se_{comp1})*Se_{comp2}	/	(prev*(1−Se_{comp1})*Se_{comp2}+(1−prev)*Sp_{comp1}*(1−Sp_{comp2}))
3	+−	prev*Se_{comp1}*(1−Se_{comp2})	/	(prev*Se_{comp1}*(1−Se_{comp2})+(1−prev)*(1−Sp_{comp1})*Sp_{comp2})
4	−−	prev*(1−Se_{comp1})*(1−Se_{comp2})	/	(prev*(1−Se_{comp1})*(1−Se_{comp2})+(1−prev)*Sp_{comp1}*Sp_{comp2})

Table 2. Distribution pattern of two component tests, mapped on a theoretical sample

Test pattern							
Component test 1	Component test 2	Probability of observing test pattern	Probability of target disease (given the test pattern)	Dichotomous class (DC)	Total[a] (n = 1,000)	Truly disease present (D1)	Truly disease absent (D0)
+	+	0.30	0.960	1	300	288	12
−	+	0.12	0.600	1	120	72	48
+	−	0.14	0.229	0	140	32	108
−	−	0.44	0.018	0	440	8	432

The table shows how a sample with a known disease prevalence is classified by the expert panel. Dichotomous class (DC) is assigned using a threshold for target disease presence probability of 50%.

[a] Although individuals are classified as disease present (DC1) or disease absent (DC0), note that not all of them are.

changed from 40% to 42%. Note that within DC1, only 85% truly has the disease present, and in DC0, 93% is truly nondiseased, hence a 15% overestimation of the number of diseased individuals in DC1, and 7% overestimation of nondiseased individuals in DC0.

Now we can construct the contingency table that we would expect to obtain for our index test when compared with the expert panel's target disease classification. DC1 consists of component test patterns ++ and −+, of which we know that 360 individuals are diseased (D1) and that the remaining 60 are nondiseased (D0). For simplicity, we will denote these diseased individuals by N(D1, DC1) and nondiseased individuals by N(D0, DC1). Given a positive index test and its sensitivity (Se_{index}) and specificity (Sp_{index}), we can calculate the number of true positives Se_{index} * N(D1, DC1) and number of false positives $(1-Sp_{index})$ *N(D0, DC1) in DC1. This can also be carried out given a negative index test, and for DC0, using the following formulas:

$$\text{Index test} + \mid \text{DC1} = Se_{index} * N(D1, DC1) + (1-Sp_{index}) * N(D0, DC1)$$
$$\text{Index test} - \mid \text{DC1} = Sp_{index} * N(D0, DC1) + (1-Se_{index}) * N(D1, DC1)$$
$$\text{Index test} + \mid \text{DC0} = Se_{index} * N(D1, DC0) + (1-Sp_{index}) * N(D0, DC0)$$
$$\text{Index test} - \mid \text{DC0} = Sp_{index} * N(D0, DC0) + (1-Se_{index}) * N(D1, DC0)$$

The resulting contingency table, which we refer to as the observed contingency table, can be found in Table 3. This table shows the shift between true disease status by a perfect (gold) reference standard and observed disease status after disease classification by the expert panel (marked by the gray arrows). The misclassification of true diseased and nondiseased individuals resulting in the observed classification is indicated by the red arrows. Consider the cell with a positive index test and disease present according to classification. In the true contingency table, this consisted of 320 individuals, whereas in the observed contingency table, this number has dropped to 300.

Values for sensitivity and specificity of the index test are calculated from the observed table, yielding 71.4% and 75.9%, respectively, which, compared with the true values of 80%, correspond to a bias of 8.6% and 4.1%, respectively. We can also derive the total proportion of misclassifications by adding the misclassifications in disease classification by the expert panel in both directions and dividing by the total number of individuals. For our example, this yields (40 + 60)/1,000 = 10% misclassification.

3. Methods of the simulation study

We investigated a series of hypothetical scenarios to study the impact of dichotomous classification of the target disease on the bias in sensitivity and specificity estimates of an index test. Each scenario consists of an expert panel with multiple component tests at their disposal, with varying accuracy (all less than 100%). The calculations as described in the preceding illustrative example were used. For further background on the methods used, we refer to the Supplementary File.

Ten different expert panel scenarios were assessed in our simulation study, described in Table 4. In the base scenario, the expert panel was provided with four component tests, each with a sensitivity and specificity of 70%, and the target disease prevalence was 20%. In other scenarios, one of the following factors was varied: the number of component tests (two, four, and eight), the diagnostic accuracy of these component tests (60%, 70%, 80%, and a combination of high and low accuracy component tests), and target disease prevalence (10%, 20%, and 40%). Threshold for dichotomizing and assigning target disease status was kept constant at 50% across all scenarios (i.e., classification to the most likely target disease status). In all scenarios, we assumed conditional independence between results of component tests.

3.1. Performance of the expert panel

Diagnostic performance of the expert panel was assessed by calculating the area under the receiver operator

Table 3. An illustrative example showing the differences in contingency tables and accuracy when comparing an index test to a gold standard (true) vs. an index test to an expert panel using two imperfect tests (observed)

	Disease classification method	Disease present according to classification	Reclassified	Disease absent according to classification	Total
Index test +	Gold standard	320	32 / 12	120	440
	Expert panel	300		140	
Index test −	Gold standard	80	8 / 48	480	560
	Expert panel	120		440	
Total	Gold standard	400	40 / 60	600	1,000
	Expert panel	420		580	

	Sensitivity	Specificity
Gold standard (true value)	320/400 = 80.0%	480/600 = 80.0%
Expert panel (observed value)	300/420 = 71.4%	440/580 = 75.9%
Bias in estimates	80.0%–71.4% = 8.6%	80.0%–75.9% = 4.1%

The shift between true and observed disease classification is given by the gray arrows.

The shift between disease present and disease absent, resulting in the observed classification, is indicated by the red arrows.

Note that of the 420 individuals classified as disease present by the expert panel, only 360 actually have the disease, and of the 580 individuals classified as disease absent by the expert panel, only 540 actually do not have the disease

Table 4. Description of expert panel scenarios

Scenario	# of component reference tests	Component reference test sensitivity	Component reference test specificity	Target disease prevalence
Low number of tests[a]	2	70%	70%	20%
Medium number of tests[a]*	4	70%	70%	20%
High number of tests[a]	8	70%	70%	20%
Low accuracy[b]	4	60%	60%	20%
Medium accuracy[b]*	4	70%	70%	20%
High accuracy[b]	4	80%	80%	20%
Mirrored accuracy[b]	4	60%–70%–80%–90%	90%–80%–70%–60%	20%
Low prevalence[c]	4	70%	70%	10%
Medium prevalence[c]*	4	70%	70%	20%
High prevalence[c]	4	70%	70%	40%

A probability of 50% was chosen as a threshold for dichotomization of the target disease.

* Base scenarios.

[a] Number of component tests altered

[b] Accuracy of component tests altered.

[c] Target disease prevalence altered.

characteristic (AUROC) and the proportion of misclassifications [21]. AUROC is a measure for overall discriminative performance of the expert panel that can be derived using probability of target disease presence as a continuous cutoff threshold. The proportion of misclassifications was calculated as the proportion of incorrect target disease classifications using the aforementioned threshold of 50% for dichotomization.

3.2. Bias in sensitivity and specificity estimates of the index test

We calculated the resulting bias in sensitivity and specificity estimates of the index test after dichotomous target disease classification by the reference standard for each of the scenarios. A comprehensive range (0–100%) of true index test sensitivity and specificity values was analyzed to assess the amount and direction of bias in each scenario. Only either index test sensitivity or specificity was varied at a time. When varying index test sensitivity, the specificity was kept constant at 80%. In a similar way, when specificity was varied, the sensitivity was fixed at 80%. Conditional independence between the index test and the component tests was assumed.

4. Results of the simulation study

4.1. Performance of the expert panel

The expected distribution of component reference test patterns and their corresponding probability of target disease presence are visualized in Fig. 1. The bars visualize the expected relative frequencies of target disease probabilities corresponding to different component test patterns. The total number of these patterns possible for a given scenario is given by two to the power of the number of component tests (i.e., $2^4 = 16$ patterns for the base scenario). One bar may contain more than one component test pattern when patterns have an equal probability of target disease presence. Target disease probability estimates toward the extremes (zero or one) are likely to yield the least incorrect classifications; almost all individuals in these patterns are likely to be either truly diseased or nondiseased; hence, forced dichotomization of the expert panel will result in minimal incorrect classifications. However, when there are patterns around the target disease dichotomization threshold (in this case 50%), and probability of observing these patterns is high, the likelihood of errors after dichotomization will increase.

As shown in the figure, when all component tests have identical accuracy, many combinations of component test patterns will have the same probability of the target disease being present. When comparing the scenarios with a low and high number of component tests, there was a higher likelihood of observing test patterns closer to the extremes for the latter, which resulted in higher discrimination (AUROC) and fewer misclassifications by the expert panel. Similar trends were observed when increasing the accuracy of the component tests. In the mirrored accuracy scenario, there was more spread in the probability of target disease presence for the various combinations of component test patterns, but overall

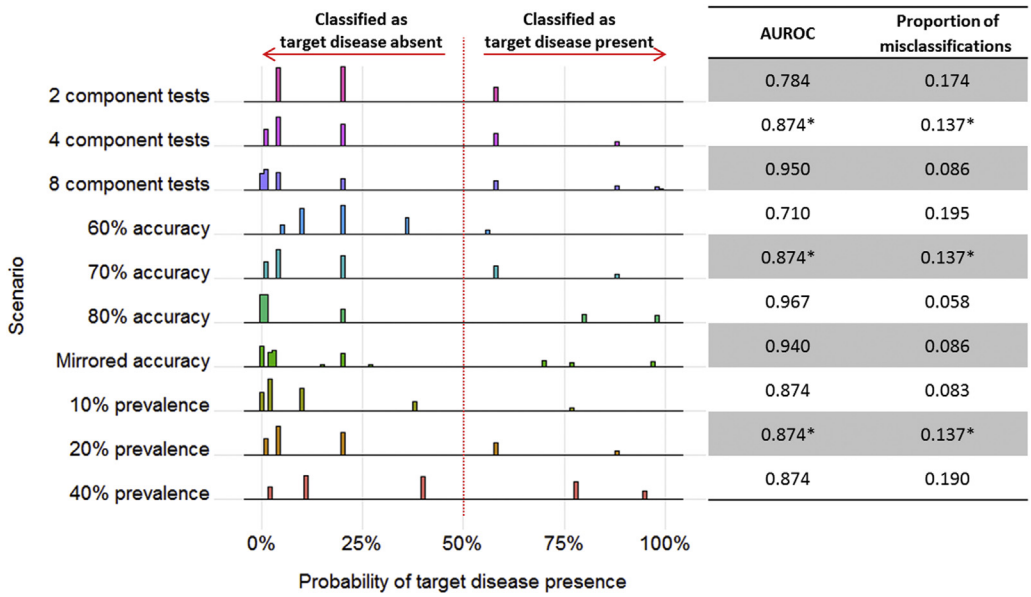

Fig. 1. Distribution of component reference test patterns and their associated probability of target disease presence, for each scenario. Proportion of misclassifications (at a threshold of 50% given by the red dotted line) and area under the receiver operator characteristic (AUROC) are given as measures of diagnostic performance. If multiple component test patterns have the same probability of disease presence, they are aggregated together in a bar. The base scenarios are marked with an asterisk. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

provided similar discriminative performance (0.940 vs. 0.967) and proportion of misclassifications (0.086 vs. 0.058) compared with the high accuracy scenario. Changes to the target disease prevalence did not affect discriminative performance; however, it did affect the expected number of misclassifications.

4.2. *Bias in sensitivity and specificity estimates of the index test*

Dichotomous target disease classifications by the expert panels in the aforementioned scenarios were used to estimate bias in sensitivity and specificity for a range of true values of an index test (Fig. 2). In all investigated scenarios, there was deviation from the reference line, indicating that in virtually all cases, there is bias in estimates of index test sensitivity and specificity. When considering the base scenario, combined with for example true values of 80% sensitivity and specificity of the index test, estimates for index test sensitivity and specificity by the expert panel were 60% and 75%, respectively, leading to an absolute bias of 20% and 5%.

The amount of bias differed across scenarios. Fig. 2A shows the shift for the low, medium, and high number of component test scenarios. A larger number of component tests resulted in a lower bias for both index test sensitivity and specificity. In a similar fashion, increasing accuracy of component tests led to less bias in estimates of sensitivity and specificity. The mirrored and high accuracy scenarios showed similar bias in estimates.

Although changes in target disease prevalence did not affect the AUROC of the reference standard (Fig. 1), it did produce irregular results in terms of bias of sensitivity and specificity of the index test. In Fig. 2C, the bias in sensitivity was highest at medium target disease prevalence, lowest at high target disease prevalence, and intermediate at the lowest target disease prevalence. This can be explained by examining the distribution of test patterns shown in Fig. 1. The difference in bias of sensitivity and specificity estimates of an index test between two scenarios was influenced by whether a component test results pattern shifted across the threshold used in the dichotomization process. For example, when looking at low and medium prevalence scenarios, there was a shift of the fourth test pattern across

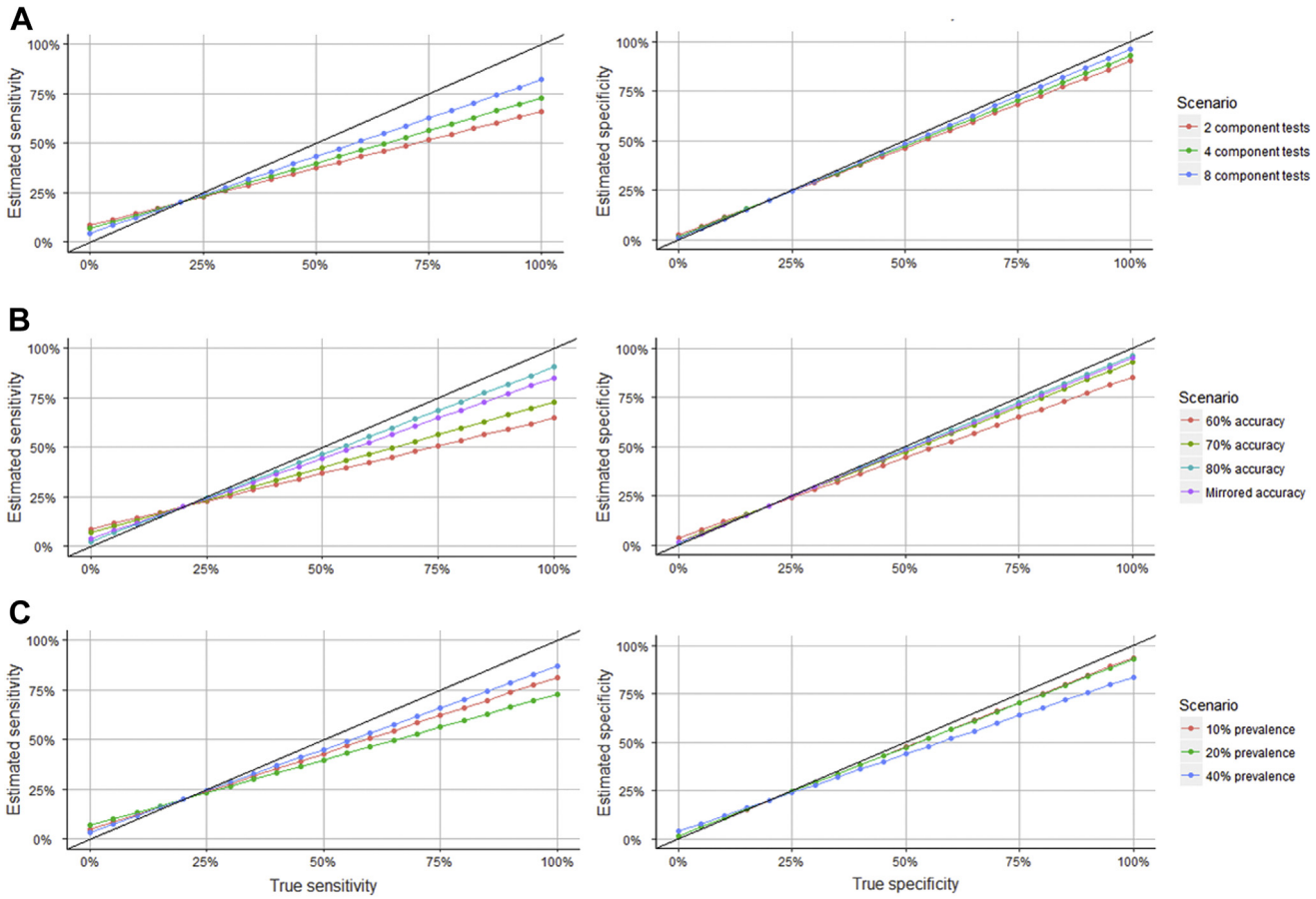

Fig. 2. Range of true values of sensitivity and specificity of a hypothetical index test and their recalculated estimates. Scenarios with different number of component reference tests (A), accuracy (B), and prevalence (C) were taken for the reference standard. A reference line is given in solid black. Dichotomization was based on a target disease probability threshold of 50%. Index test specificity was fixed at 80% when calculating sensitivity, and vice versa.

the 50% threshold. Individuals with that test pattern result were suddenly all classified as disease being absent in the low-prevalence scenario, and all as disease being present in the medium prevalence scenario. As a consequence, there was a strong increase in bias of sensitivity estimates, whereas the effect on bias of specificity estimates was limited.

5. Discussion

Forcing expert panels to dichotomize target disease classification leads to both target disease misclassification and biased accuracy estimates of the index test under study, even when individuals are consistently classified to their most likely target disease status. A series of scenarios were assessed in which an expert panel was given a set of component reference tests with varying characteristics combined with a range of true accuracy values for the index test. Virtually all scenarios lead to biased index test accuracy estimates. Increasing the number and/or accuracy of component reference tests reduced bias in the index test accuracy estimates. Varying target disease prevalence led to irregular shifts in bias of index test accuracy.

The scenarios that were investigated demonstrated a structural underestimation of index test sensitivity and/or specificity when (realistic) true values of at least 50% for both parameters were considered. However, it would be an error to assume that index test accuracy will always be underestimated when expert panels are used as a reference standard in diagnostic studies. In particular, the index test results might be correlated (conditionally dependent) for a given true disease status, which might lead to overestimation rather than underestimation of sensitivity and/or specificity of the index test. In addition, in case of conditional dependence between component reference test results, adding more component tests may not always improve estimation of the accuracy of the index test [14].

When looking at the distribution of the probability of target disease presence for different component test patterns (Fig. 1), one might anticipate that a symmetrical distribution (i.e., equal distributions left and right of the threshold) will cancel out any target disease misclassifications made by a reference standard, which should then consequently reduce bias in accuracy estimates of the index test. When we simulated a scenario with such a symmetrical distribution, bias in estimated index test sensitivity and specificity were equal across the range of true values; however, more research is required to investigate whether this minimalizes bias in sensitivity and specificity of the index test.

One tempting option in diagnostic studies would be to exclude individuals where there is significant uncertainty about the true disease status, as these have the highest probability of leading to erroneous target disease classification by the expert panel. However, this is ill-advised. Excluding cases in which there is uncertainty about the true disease status (i.e., close to the threshold) would mean the accuracy of the index test would only be generalizable to the assessment of the ''easy'' cases that have a high probability of either having or not having the target disease. This obviously does not represent the true target population of the index test; hence, such study patient exclusion will yield a distorted and too optimistic accuracy of the index test. Similar issues have been described for diagnostic case−control studies [22−24].

Earlier studies have demonstrated similar effects on estimates of sensitivity and specificity of index tests when composite reference standards based on explicit decision rules were used [14,25]. Expert panels as reference standards deal with similar issues as these composite reference standards, resulting in biased index test accuracy estimates. However, unlike composite reference standards with explicit decision rules, we studied the effect of target disease dichotomization based on the probability of target disease presence, which is commonly ignored when developing a composite reference standard. A recent article expressed further concerns regarding such types of composite reference standards, and suggested alternatives such as latent class models to take into account uncertainty surrounding target disease classification [15].

An alternative approach to minimize bias from dichotomous classification of target disease status would be to allow for probabilistic target disease estimates on a continuous or ordinal scale, which have already been applied in a few diagnostic settings [26,27]. Taking such probabilistic estimates of target disease presence is currently seldom being applied in studies exploiting expert panels as a reference standard, however have been described in the context of record linkage [28,29]. Some authors have suggested obtaining ordinal target disease classes between the traditional disease present and absent options, such as ''possible disease'' and ''intermediate classes'' [30,31]. Others have suggested using methods such as diagnostic probability functions based on expert diagnosis to obtain target disease probabilities [27]. Although, it has been emphasized that eliciting expert judgments on disease status is a complex task [32].

To fully appreciate the findings of this article, there are some limitations that should be considered. First, in our simulations, we have only considered dichotomous component reference tests, whereas in practice, some test results may produce continuous outcomes. Unless these continuous tests can be used to perfectly separate individuals with and without the target disease, uncertainty in target disease classifications will remain present. Therefore, bias in index test accuracy estimates after dichotomization of target disease classification based on continuous diagnostic component reference tests, is also to be anticipated.

Secondly, we have not included conditional dependence between component reference tests nor between component tests and the index test. Conditional dependence is likely to

be present in real-life situations, for instance, because tests are likely to make fewer errors in more severe cases compared with less severe cases [33]. We anticipate that similar problems as observed would occur for test results that are conditionally dependent. The exact influence of dependent test results may be a complicated interplay between the mechanism of the dependence between the tests, which may obviously vary between settings, the accuracy of the component tests and index test, and the prevalence of the disease [14]. While our results may be viewed as a simplification, the fact that the bias occurs even in the simplest situations should already be of great concern.

Finally, we have assumed that the expert panel is able to correctly estimate the target disease probability for all individuals, and that these individuals are consistently classified to the target disease status with the highest probability. In diagnostic research, this may not always be realistic, especially when target disease probability estimates of individuals are centered around the threshold for dichotomous classification. Thus, we may expect even more target disease classification errors when not all subjects are classified to the target disease status with the highest probability.

Our findings are not only applicable to expert panels serving as a reference standard in diagnostic studies, but also to other situations in which a dichotomous outcome classification is used, and uncertainty is not taken into account. Composite reference standards in diagnostic research [34,35], adjudication committees used to classify endpoints in intervention or prognostic studies [36], and probabilistic medical record linkage [28,37,38], frequently force dichotomization from their respective reference standards. As a result, similar biases may occur.

We conclude that dichotomizing target disease classification by a reference standard based on multiple imperfect component tests, such as a panel diagnosis, leads to biased accuracy estimates of the index test in a simulation study. The direction and magnitude of these biases were found to depend on the combination of the number of component reference tests, their accuracy, and the target disease prevalence. The bias found in this simulation study may not reflect the true bias in an empirical setting, as more complex interactions, such as conditional dependence and misclassification by expert panels (e.g., classifying an individual with a low probability of disease, as target disease present) may be at play. To potentially reduce these biases, alternatives to dichotomous classification of target disease by the reference standard should be sought after, such as obtaining target disease probability estimates per individual from the expert panel, or via a latent class analysis [39]. Researchers involved in diagnostic studies that employ expert panels as a reference standard should be wary that solely asking for presence or absence of the target disease will limit the ability of unbiased estimation of index test accuracy. Performance of novel diagnostic tests needs to be established accurately in diagnostic research, and it should not have to suffer from the imperfectness of reference standard that it is being compared with.

CRediT authorship contribution statement

Kevin Jenniskens: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing - original draft, Writing - review & editing. **Christiana A. Naaktgeboren:** Conceptualization, Methodology, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Johannes B. Reitsma:** Conceptualization, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Lotty Hooft:** Writing - review & editing. **Karel G.M. Moons:** Conceptualization, Writing - review & editing. **Maarten van Smeden:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.03.002.

References

[1] Bachmann LM, Juni P, Reichenbach S, Ziswiler HR, Kessels AG, Vogelin E. Consequences of different diagnostic "gold standards" in test accuracy research: carpal Tunnel Syndrome as an example. Int J Epidemiol 2005;34:953−5.

[2] Reference standard. Mosby's medical dictionary. 8th ed. 2009. Available at https://medical-dictionary.thefreedictionary.com/Reference+Standard. Accessed May 2, 2018.

[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;62:797−806.

[4] 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:W1−12.

[5] Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A Review Methods. Health Technol Assess 2007;11(50):iii. ix−51.

[6] Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med 2013;10(10):e1001531.

[7] Bertens LC, van Mourik Y, Rutten FH, Cramer MJ, Lammers JW, Hoes AW, et al. Staged decision making was an attractive alternative to a plenary approach in panel diagnosis as reference standard. J Clin Epidemiol 2015;68:418−25.

[8] Bossuyt PM. Interpreting diagnostic test accuracy studies. Semin Hematol 2008;45(3):189−95.

[9] Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. BMJ 2002;324:477−80.

[10] Trikalinos TA, Balion CM. Chapter 9: options for summarizing medical test performance in the absence of a "gold standard". J Gen Intern Med 2012;27:S67−75.

[11] Miller DP, O'Shaughnessy KF, Wood SA, Castellino RA, editors. Gold standards and expert panels: a pulmonary nodule case study

with challenges and solutions. Medical Imaging 2004. San Diego, CA: SPIE; 2004.
[12] Handels RL, Wolfs CA, Aalten P, Bossuyt PM, Joore MA, Leentjens AF, et al. Optimizing the use of expert panel reference diagnoses in diagnostic studies of multidimensional syndromes. BMC Neurol 2014;14:190.
[13] Thomeer M, Demedts M, Behr J, Buhl R, Costabel U, Flower CD, et al. Multidisciplinary interobserver agreement in the diagnosis of idiopathic pulmonary fibrosis. Eur Respir J 2008;31(3):585−91.
[14] Schiller I, van Smeden M, Hadgu A, Libman M, Reitsma JB, Dendukuri N. Bias due to composite reference standards in diagnostic accuracy studies. Stat Med 2016;35:1454−70.
[15] Dendukuri N, Schiller I, de Groot J, Libman M, Moons K, Reitsma J, et al. Concerns about composite reference standards in diagnostic research. BMJ 2018;360:j5779.
[16] Johnson S, Hollis C, Hennessy E, Kochhar P, Wolke D, Marlow N. Screening for autism in preterm children: diagnostic utility of the Social Communication Questionnaire. Arch Dis Child 2011;96:73−7.
[17] Brugha TS, McManus S, Smith J, Scott FJ, Meltzer H, Purdon S, et al. Validating two survey methods for identifying cases of autism spectrum disorder among adults in the community. Psychol Med 2012;42:647−56.
[18] Fryback DG. Bayes' theorem and conditional nonindependence of data in medical diagnosis. Comput Biomed Res 1978;11(5):423−34.
[19] Simon D, Boring JR III. Sensitivity, specificity, and predictive value. In: Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations. Boston: Butterworths; 1990.
[20] Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol 2008;56:45−50.
[21] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29−36.
[22] Song JW, Chung KC. Observational studies: cohort and case-control studies. Plast Reconstr Surg 2010;126:2234−42.
[23] Kopec JA, Esdaile JM. Bias in case-control studies. A review. J Epidemiol Community Health 1990;44:179−86.
[24] Thomas SV, Suresh K, Suresh G. Design and data analysis case-controlled study in clinical research. Ann Indian Acad Neurol 2013;16(4):483−7.
[25] Walter SD, Macaskill P, Lord SJ, Irwig L. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. Stat Med 2012;31:1129−38.
[26] van Houten CB, de Groot JA, Klein A, Srugo I, Chistyakov I, de Waal W, et al. A host-protein based assay to differentiate between bacterial and viral infections in preschool children (OPPORTUNITY): a double-blind, multicentre, validation study. Lancet Infect Dis 2017;17:431−40.
[27] Steurer J, Held U, Miettinen OS. Diagnostic probability function for acute coronary heart disease garnered from experts' tacit knowledge. J Clin Epidemiol 2013;66:1289−95.
[28] Hof MH, Zwinderman AH. Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. Stat Med 2012;31:4231−42.
[29] Hof MH, Zwinderman AH. A mixture model for the analysis of data derived from record linkage. Stat Med 2015;34:74−92.
[30] van Mourik Y, Bertens LC, Cramer MJ, Lammers JW, Reitsma JB, Moons KG, et al. Unrecognized heart failure and chronic obstructive pulmonary disease (COPD) in frail elderly detected through a near-home targeted screening strategy. J Am Board Fam Med 2014;27(6):811−21.
[31] Poldervaart JM, Reitsma JB, Backus BE, Koffijberg H, Veldkamp RF, Ten Haaf ME, et al. Effect of using the HEART score in patients with chest pain in the emergency department: a stepped-wedge, cluster randomized trial. Ann Intern Med 2017;166:689−97.
[32] O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. Uncertain judgements: eliciting experts' probabilities. Chichester: John Wiley & Sons; 2006.
[33] Brenner H. How independent are multiple 'independent' diagnostic classifications? Stat Med 1996;15:1377−86.
[34] Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Stat Med 1999;18:2987−3003.
[35] Naaktgeboren CA, Bertens LC, van Smeden M, de Groot JA, Moons KG, Reitsma JB. Value of composite reference standards in diagnostic research. BMJ 2013;347:f5605.
[36] Sepehrvand N, Zheng Y, Armstrong PW, Welsh R, Goodman SG, Tymchak W, et al. Alignment of site versus adjudication committee-based diagnosis with patient outcomes: insights from the providing rapid out of hospital acute cardiovascular treatment 3 trial. Clin trials 2016;13:140−8.
[37] Scheuren F, Winkler WE. Regression analysis of data files that are computer matched. Surv Methodol 1993;19(1):39−58.
[38] Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. Int J Epidemiol 2016;45:954−64.
[39] van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KG, de Groot JA. Latent class models in diagnostic studies when there is no reference standard–a systematic review. Am J Epidemiol 2014;179:423−31.