Statistics
in Medicine

Research Article

Received 27 February 2015, Accepted 19 October 2015 Published online 10 November 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6803

Bias due to composite reference standards in diagnostic accuracy studies

Ian Schiller,[a] Maarten van Smeden,[b] Alula Hadgu,[c] Michael Libman,[d] Johannes B. Reitsma[b] and Nandini Dendukuri[a]*†

Composite reference standards (CRSs) have been advocated in diagnostic accuracy studies in the absence of a perfect reference standard. The rationale is that combining results of multiple imperfect tests leads to a more accurate reference than any one test in isolation. Focusing on a CRS that classifies subjects as disease positive if at least one component test is positive, we derive algebraic expressions for sensitivity and specificity of this CRS, sensitivity and specificity of a new (index) test compared with this CRS, as well as the CRS-based prevalence. We use as a motivating example the problem of evaluating a new test for *Chlamydia trachomatis*, an asymptomatic disease for which no gold-standard test exists. As the number of component tests increases, sensitivity of this CRS increases at the expense specificity, unless all tests have perfect specificity. Therefore, such a CRS can lead to significantly biased accuracy estimates of the index test. The bias depends on disease prevalence and accuracy of the CRS. Further, conditional dependence between the CRS and index test can lead to over-estimation of index test accuracy estimates. This commonly-used CRS combines results from multiple imperfect tests in a way that ignores information and therefore is not guaranteed to improve over a single imperfect reference unless each component test has perfect specificity, and the CRS is conditionally independent of the index test. When these conditions are not met, as in the case of *C. trachomatis* testing, more realistic statistical models should be researched instead of relying on such CRSs. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: imperfect reference; composite; sensitivity; specificity; conditional dependence

1. Introduction

For many diseases, there is no gold standard diagnostic test having both perfect sensitivity and specificity. Obtaining a definitive diagnosis for each subject in a diagnostic test evaluation study or disease prevalence study therefore becomes challenging [1]. Disease classification based on any imperfect single reference test will lead to biased inferences [2–4].

Consider the problem of evaluating a new test for *Chlamydia trachomatis* (*C. trachomatis*), an asymptomatic, infectious disease with harmful medical consequences if missed, and potentially serious social consequences if falsely diagnosed. The reference test of choice was previously cell culture, which is thought to have near-perfect specificity but only moderate sensitivity [5]. More recently, developed nucleic acid amplification tests (NAATs) are considered more sensitive than culture, but not as specific.

To reduce the misclassification of *C. trachomatis* status, a number of studies have used a composite reference standard (CRS) for evaluating a new test (or index test) [6–11]. For example, Alonzo *et al.* [6] defined a CRS based on two imperfect tests-cell culture and polymerase chain reaction (PCR). Subjects were classified as *C. trachomatis*-positive if positive on at least one of the two component tests and as negative otherwise. This CRS is based on what we will refer to as an OR decision rule (i.e., logical disjunction or 'any positive' rule) according to which a positive result on one or more component tests is defined as having the disease of interest. Another common way to define a CRS is the AND decision

[a] *Division of Clinical Epidemiology, McGill University Health Centre, Montreal, Canada*
[b] *Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht, The Netherlands*
[c] *Division of STD Prevention, Centers for Disease Control, Atlanta, GA 30329-4027, U.S.A.*
[d] *Division of Infectious Diseases, McGill University Health Centre, Montreal, Canada*
**Correspondence to: Nandini Dendukuri, Division of Clinical Epidemiology, McGill University Health Centre, 687 Pine Avenue W R4.09, Montreal, Quebec H3A 1A1, Canada.*
† *E-mail: nandini.dendukuri@mcgill.ca*

Copyright © 2015 John Wiley & Sons, Ltd.

rule (i.e., logical conjunction or 'all positive' rule) according to which a positive result on all component tests is required to classify a subject as having the disease. Besides applications in *C. trachomatis*, CRSs are widely used, for example, for extra-pulmonary tuberculosis, community-acquired pneumonia, and pertussis [12–14].

The CRS is appealing because it provides a simple rule to assign a final 'diagnosis' to each study subject. It is also described as being 'unaffected by' or 'independent of' the test under evaluation because the test under evaluation is not used in defining the final diagnosis [6, 8]. However, although the CRS aims to reduce misclassifications of the disease status, it will not eliminate them [15]. Nor are the component tests necessarily independent of the index test [16]. For *C. trachomatis*, the aim is to reduce the number of false negatives by increasing the number of component tests in combination with the OR-rule. The prospect of reducing misclassification by increasing the number of component tests has led to a diversity of CRSs encountered in practice – some having as few as two component tests [6, 17], others as many as 9 [10], while the Food and Drug Administration (FDA) draft guidance requires four [18].

So far, little attention has been paid to study the accuracy of the CRS itself and to how its accuracy is affected by its composition [16, 19]. Another aspect that has not received much attention is the impact of dependence between the errors of the component and index tests. Alternative statistical methods, particularly latent class analysis, have been the subject of much scrutiny with regard to the impact of conditional dependence, and studies have shown that ignoring conditional dependence can result in over-estimation of accuracy of the index test [20–23]. Dependence of errors may occur, for example, if there exists a spectrum of disease severity with both component and index tests being more likely to detect more severe cases and to miss less severe cases [24, 25] or when component and index tests are based on the same technology. In the approach currently used by the FDA to evaluate new NAATs for *C. trachomatis* infection, a new test is typically compared against a CRS made up of other NAATs based on the same biological mechanism [18, 26]. It is highly likely that there is a conditional dependence between the CRS and the test under evaluation in this setting.

In this article, we aim to bridge these gaps by elucidating the workings of the CRS. We will focus on the OR definition of the CRS as this applies to most *C. trachomatis* settings, where the specificity of the component tests are considered to be relatively high and there is an interest in improving the overall sensitivity of the CRS beyond that of any single component test. In Section 2, we describe a motivating example of evaluating a new test for *C. trachomatis*. In Section 3, we derive algebraic expressions for the accuracy of the CRS, as well as expressions for the index test's accuracy based on the CRS. In Sections 4 and 5, we examine the behavior of these parameters in several simulated scenarios. In section 6, we briefly investigate CRSs defined by other decision rules. We conclude with a Discussion.

2. Estimating the accuracy of a test for *C. trachomatis* infection

To illustrate application of a CRS in practice, we use data from a diagnostic accuracy study of *C. trachomatis* tests [8]. Data are available on 743 asymptomatic women (see Supporting Information) on three types of tests: (i) four NAATs: PCR and ligase chain reaction (LCR) tests each carried out on both cervical (PCRC, LCRC) and urine specimens (PCRU, LCRU); (ii) two culture tests one each on cervical (CULC) and urethral specimens (CULU); and (iii) a DNA hybridization test (DNAP).

Although based on different mechanisms, all tests are designed to have high specificity [27]. NAATs are designed to detect *C. trachomatis* DNA at a low organism load and are therefore considered most sensitive [5]. However, these tests cannot distinguish between DNA from viable and non-viable organisms contributing to less than perfect specificity. In comparison, cell culture and DNAP tests have lower sensitivity [5]. The cervical culture test is believed to have near perfect specificity [8]. We use one NAAT(PCRU), which was new at the time of the data collection, as the index test.

To evaluate the robustness of CRS-based accuracy estimates, we considered different CRSs that could reasonably be defined with the available data. To examine the impact of increasing the number of component tests, we defined a sequence of reference standards starting with a single test (CULC) followed by CRS1 = CULC+ OR CULU+, CRS2= CRS1+ OR DNAP+, CRS3=CRS2+ OR LCRC+ and CRS4=CRS3+ OR LCRU+.

When the reference standard was CULC, estimated sensitivity of the index test PCRU was 0.83 ($95\% CI, 0.75 - 0.89$), but it decreased to 0.78 ($95\% CI, 0.70 - 0.83$) using a CRS with five component tests (Table I). Simultaneously, the estimated specificity of PCRU ranged from 0.94 ($95\% CI, 0.92 - 0.96$) based on CULC to 0.99 ($95\% CI, 0.98 - 0.99$) based on five component tests. Our example shows that estimates of sensitivity and specificity of the index test are clearly sensitive to composition of the reference

Copyright © 2015 John Wiley & Sons, Ltd.

Table I. Impact of increasing number of tests in the CRS on estimated sensitivity ($\hat{S}^*_{PCRU}$) and estimated specificity ($\hat{C}^*_{PCRU}$) of the PCRU test. (For each CRS component tests used in OR decision rule appear in brackets.)

	Estimate (95% Confidence Interval)	
Reference standard	$\hat{S}^*_{PCRU}$	$\hat{C}^*_{PCRU}$
CULC+	0.833 (0.752 – 0.892)	0.943 (0.923 – 0.959)
CRS1 (CULC+ or CULU+)	0.825 (0.747 – 0.883)	0.957 (0.938 – 0.970)
CRS2 (CRS1+ or DNAP+)	0.810 (0.732 – 0.869)	0.961 (0.943 – 0.974)
CRS3 (CRS2+ or LCRC+)	0.804 (0.705 – 0.838)	0.983 (0.969 – 0.991)
CRS4 (CRS3+ or LCRU+)	0.776 (0.704 – 0.834)	0.992 (0.980 – 0.996)

standard. This has also been noted by others, for example, in a systematic review of tests for *C trachomatis*, Cook *et al.* [28] found that those studies that had used a CRS with more component tests reported lower sensitivity and higher specificity. To understand how the composition of the CRS affects its accuracy and the estimated accuracy of the index test, we derive analytical expressions for the properties of the CRS as a function of its component tests in the following section.

3. Accuracy of composite reference standards and accuracy of index test with respect to composite reference standards

Let $T_i = (T_{i1}, \dots, T_{iP})$ denote the vector of observed diagnostic test results for the i^{th} subject. T_{ij} takes values one if positive and zero if negative on test j, and $Pr(T_i)$ denotes the probability function of observing T_i. The first $(P-1)$ tests are used to define the CRS. The P^{th} test is an index test to be evaluated. The CRS based on the OR-rule is formally defined as

$$CRS_i = I(T_{i1}, \dots, T_{iP-1}) = \begin{cases} 1 & \text{if } max(T_{i1}, \dots, T_{iP-1}) = 1 \\ 0 & \text{if } max(T_{i1}, \dots, T_{iP-1}) = 0, \end{cases} \tag{1}$$

where I is an indicator function and CRS_i=1 implies the i^{th} subject is classified as having the disease. For brevity, we drop the subscript i from the remainder of the presentation.

In the development later, we examine two scenarios: (i) the P tests (i.e., both component and index) are independent conditional on the true (unobserved) disease status; (ii) the P tests are dependent conditional on the true disease status.

3.1. *Case when tests are conditionally independent*

Assume that all P tests are stochastically independent conditional on the target disease status. The target disease status, denoted by D, takes the value $d = 0$ if the target disease is absent, and $d = 1$ if it is present. The probability function of each combination of test results can be written as

$$Pr(T) = \sum_{d=0}^{1} Pr(T_1, \dots, T_P|D = d)Pr(D = d) = \sum_{d=0}^{1} Pr(D = d)\prod_{j=1}^{P} Pr(T_j|D = d). \tag{2}$$

The sensitivity of the CRS based on the OR-rule is given by

$$\begin{aligned} S_{CRS} &= Pr(CRS = 1|D = 1) = Pr(max(T_1, \dots, T_{P-1}) = 1|D = 1) \\ &= 1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 1|D = 1)) = 1 - \prod_{j=1}^{P-1}(1 - S_j), \end{aligned} \tag{3}$$

where S_j denotes the true sensitivity of the j^{th} component test. The specificity of the CRS is given by

$$\begin{aligned} C_{CRS} &= Pr(CRS = 0|D = 0) = Pr(max(T_1, \dots, T_{P-1}) = 0|D = 0) \\ &= \prod_{j=1}^{P-1} Pr(T_j = 0|D = 0) = \prod_{j=1}^{P-1} C_j, \end{aligned} \tag{4}$$

Copyright © 2015 John Wiley & Sons, Ltd.

where C_j denotes the true specificity of the j^{th} component test.

A perfect CRS would have zero probability of misclassification, that is: $Pr(CRS = 1|D = 1) = Pr(CRS = 0|D = 0) = 1$. From Equations (3) and (4), it follows that for a CRS based on a finite number of component tests, under the assumption of conditional independence of all component tests, the OR-rule leads to a perfect CRS if and only if the following conditions are satisfied:

(1) $C_1 = C_2 = \ldots = C_{(P-1)} = 1$,
(2) $S_j = 1$ for at least one $j = 1, \ldots, P-1$.

Put in words, the conditions given previously imply that for a CRS to be perfect it should include at least one component test with perfect sensitivity and specificity (i.e., is a gold standard which would then make the use of the CRS unnecessary), and all other component tests should at least have perfect specificity. Clearly, these conditions will not be met in practice. It is therefore of interest to see how the use of the CRS influences the estimated accuracy of the index tests.

Let the sensitivity of index test T_P with respect to the CRS be denoted by S^*_P. It can be shown that,

$$\begin{aligned} S^*_P &= Pr(T_P = 1|CRS = 1) \\ &= \frac{\sum_{d=0}^{1} Pr(T_P = 1|D = d)\left\{1 - \prod_{j=1}^{P-1}\left(1 - Pr(T_j = 1|D = d)\right)\right\} Pr(D = d)}{\sum_{d=0}^{1}\left\{1 - \prod_{j=1}^{P-1}\left(1 - Pr(T_j = 1|D = d)\right)\right\} Pr(D = d)}, \end{aligned} \tag{5}$$

illustrating that CRS based sensitivity of index test is a function of the sensitivity and specificity of the component tests and the true disease prevalence. From Equation (5), it can be seen that the sensitivity of the index test with respect to the CRS will be exactly equal to the true sensitivity of the index test ($S^*_P = S_P$) if any of the following conditions are satisfied:

(1) $Pr(D = 1) = 1$
(2) $C_1 = C_2 = \cdots = C_{(P-1)} = 1$
(3) $S_P = 1 - C_P$.

The specificity of index test P with respect to the CRS is given by

$$\begin{aligned} C^*_P &= Pr(T_P = 0|CRS = 0) \\ &= \frac{\sum_{d=0}^{1} Pr(T_P = 0|D = d)Pr(D = d)\prod_{j=1}^{P-1} Pr(T_j = 0|D = d)}{\sum_{d=0}^{1} Pr(D = d)\prod_{j=1}^{P-1} Pr(T_j = 0|D = d)}, \end{aligned} \tag{6}$$

and thus also relies on the accuracy of the component tests and the true disease prevalence. It follows directly from Equation (6) that under conditional independence of all P tests, the OR-rule leads to $C^*_P = C_P$ if any of the following conditions are satisfied:

(1) $Pr(D = 0) = 1$
(2) $S_1 = S_2 = \cdots = S_{(P-1)} = 1$
(3) $C_P = 1 - S_P$.

Considered in the context of our example on *C. trachomatis*, the Conditions 1 and 3 above for unbiased sensitivity or unbiased specificity are not relevant. A diagnostic study is unlikely to be designed in a population where the disease is either present or absent in 100% of individuals, nor is it realistic that the index test has no discriminatory value (i.e., $S_P + C_P = 1$). Under conditional independence, an unbiased estimate of the sensitivity (specificity respectively) of an index test would be obtained only if all specificities (sensitivities respectively) of $P - 1$ component tests equal one. This means that even in the ideal scenario where all component tests have perfect specificity leading to S^*_P being unbiased, C^*_P will be biased. For the case of *C. trachomatis*, while it is true that cell culture is believed to have near perfect specificity, the condition that all component tests have perfect specificity will not be met. We will study the consequences of having high, although not perfect, specificities on S^*_P and C^*_P in Section 4.

The estimated disease prevalence with respect to the CRS is defined by

$$Pr(CRS = 1) = Pr(D = 1)\left(1 - \prod_{j=1}^{P-1}\left(1 - S_j\right)\right) + Pr(D = 0)\left(1 - \prod_{j=1}^{P-1} C_j\right). \tag{7}$$

Copyright © 2015 John Wiley & Sons, Ltd.

3.2. Case when tests are conditionally dependent

Conditional dependence between the P observed tests would arise due to their association with a variable besides the disease status D. Here, we will study the situation where conditional dependence is due to two dichotomous random variables Z_0 and Z_1 – the variable Z_1 causing dependence among subjects where $D = 1$ and the variable Z_0 causing dependence among subjects where $D = 0$. Conditional dependence between component and index tests among true disease positives would imply they both have imperfect sensitivity. Likewise for specificity.

In the case of our motivating example on *C. trachomatis* tests, Z_1 could be the severity of the disease status (or organism load) among those with the infection, while Z_0 could represent the presence of $Chlamydia$ DNA among subjects who had a recent infection but are currently disease negative. The three dichotomous variables Z_1, Z_0, and D define between them four possible classes of subjects. We will denote these classes by the following: $L_1 = (Z_1 = 1, D = 1), L_2 = (Z_1 = 0, D = 1), L_3 = (Z_0 = 1, D = 0)$ and $L_4 = (Z_0 = 0, D = 0)$. The joint probability function of the observed test results can be defined as follows:

$$\begin{aligned} Pr(T) &= \sum_{Z_1=0}^{1} Pr(T_1, \dots, T_P|Z_1, D = 1)Pr(Z_1, D = 1) \\ &\quad + \sum_{Z_0=0}^{1} Pr(T_1, \dots, T_P|Z_0, D = 0)Pr(Z_0, D = 0), \\ &= \sum_{k=1}^{4} Pr(T_1, \dots, T_P|L_k)Pr(L_k) = \sum_{k=1}^{4} Pr(L_k) \prod_{p=1}^{P} Pr(T_p|L_k), \end{aligned} \tag{8}$$

assuming conditional independence between all P tests within the four classes. Notice that when the P tests are independent conditional on the target disease status $Pr(T_p|L_1) = Pr(T_p|L_2)$ and $Pr(T_p|L_3) = Pr(T_p|L_4)$.

The sensitivity and specificity of the j^{th} test with respect to the true disease status are given by

$$\begin{aligned} S_j &= Pr(T_j = 1|D = 1) = \frac{Pr(T_j = 1|L_1)Pr(L_1) + Pr(T_j = 1|L_2)Pr(L_2)}{Pr(L_1) + Pr(L_2)}, \text{ and} \\ C_j &= Pr(T_j = 0|D = 0) = \frac{Pr(T_j = 0|L_3)Pr(L_3) + Pr(T_j = 0|L_4)Pr(L_4)}{Pr(L_3) + Pr(L_4)}. \end{aligned} \tag{9}$$

The sensitivity and specificity of the CRS are given by the following:

$$\begin{aligned} S_{CRS} &= \frac{\sum_{k=1}^{2} Pr(L_k)\left(1 - \prod_{j=1}^{P-1}\left(1 - Pr\left(T_j = 1|L_k\right)\right)\right)}{\sum_{k=1}^{2} Pr(L_k)}, \text{ and} \\ C_{CRS} &= \frac{\sum_{k=3}^{4} Pr(L_k)\prod_{j=1}^{P-1}\left(1 - Pr\left(T_j = 1|L_k\right)\right)}{\sum_{k=3}^{4} Pr(L_k)}. \end{aligned} \tag{10}$$

The sensitivity and specificity of index test P with respect to the CRS are given by

$$S_P^* = \frac{\sum_{k=1}^{4} Pr\left(T_P = 1|L_k\right)Pr(L_k)\left\{1 - \prod_{j=1}^{P-1}\left(1 - Pr\left(T_j = 1|L_k\right)\right)\right\}}{\sum_{k=1}^{4}\left\{1 - \prod_{j=1}^{P-1}\left(1 - Pr\left(T_j = 1|L_k\right)\right)\right\}Pr(L_k)}, \tag{11}$$

$$C_P^* = \frac{\sum_{k=1}^{4} Pr\left(T_P = 0|L_k\right)Pr(L_k)\prod_{j=1}^{P-1} Pr\left(T_j = 0|L_k\right)}{\sum_{k=1}^{4} Pr(L_k)\prod_{j=1}^{P-1} Pr\left(T_j = 0|L_k\right)}. \tag{12}$$

Copyright © 2015 John Wiley & Sons, Ltd.

4. Examining the sensitivity and specificity of the composite reference standards

Using the expressions from Section 3, we study the effect of various factors on S_{CRS} and C_{CRS}. We consider the true sensitivity, specificity, and disease prevalence values in the following ranges: $S_j \in (0.30, 0.90)$, $C_j \in (0.90, 0.99)$ and $Pr(D = 1) \in (0.05, 0.30)$. These values are motivated by those expected for testing of *C. trachomatis* and other infectious diseases, such as tuberculosis or pneumonia, where sensitivities can range from low to high depending on the type of test, specificities are generally high, and disease prevalence is low in the general population but can be higher in high-risk sub-groups. For ease of illustration, we assume that all $P - 1$ component tests used to define the CRS have identical sensitivity and specificity, that is, $S_j = S, C_j = C, j = 1, \ldots, P - 1$. The number of component tests are varied from 2 to 10.

We also examine the case where the observations are generated under the model specified by Equation (8), where the P tests are conditionally dependent. The prevalence of the four classes are set at $Pr(L_1) = 0.05$, $Pr(L_2) = 0.05$, $Pr(L_3) = 0.05$, and $Pr(L_4) = 0.85$. The following constraints are applied without loss of generality: $Pr(T_j = 1|L_1) > Pr(T_j = 1|L_2)$ and $Pr(T_j = 1|L_3) > Pr(T_j = 1|L_4)$, $j = 1, \ldots P - 1$.

4.1. Accuracy of composite reference standards under conditional independence

From Equation (3), it follows that S_{CRS} depends only on the sensitivities (S) of the component tests. As the number of component tests increases, the sensitivity of the CRS itself increases. A simple intuitive explanation for this relation is that as the number of component tests increases the number of patients classified with a 'positive' diagnosis increases, and therefore, the probability of correctly classifying at least one target disease positive subject increases. Table II shows that combining two component tests both with sensitivity $S = 0.6$ will lead to a composite reference standard having sensitivity 0.84, that is, the probability that at least one of the two tests correctly diagnose a disease positive subject is 0.84. By simply adding a third component test with sensitivity 0.6, the CRS's sensitivity reaches 0.94. In theory if we let the number of component tests tend to infinity, the sensitivity would approach one (see Equation (3)).

Similarly, from Equation (4), it can be seen that the specificity of the CRS is completely determined by the specificities (C) of the component tests. The specificity of the CRS is a decreasing function of the number of tests. Again, the intuitive explanation is that the probability of all component tests having a negative outcome for a given target disease negative subject becomes lower as the number of tests increases. As shown in Table II, two tests with $C = 0.95$ specificity would together form a CRS with specificity 0.90. Adding a third test with the same properties ($S = 0.6, C = 0.95$) would drive the CRS's specificity down to 0.86. In theory, adding an infinite number of tests would reduce the specificity of the CRS toward zero (see Equation (4)). Thus, we observe that one cannot expect to improve both CRS sensitivity and specificity simultaneously, unless $C = 1$ for all component tests.

Table II. Relation between number of component tests and the accuracy of a CRS under settings of conditional independence and conditional dependence ((a) strong, (b) weak).

			Conditional dependence			
	Conditional independence		Setting (a)		Setting (b)	
$P-1$	S_{CRS}	C_{CRS}	S_{CRS}	C_{CRS}	S_{CRS}	C_{CRS}
2	0.840	0.903	0.800	0.930	0.840	0.903
3	0.936	0.857	0.888	0.917	0.936	0.857
4	0.974	0.815	0.934	0.908	0.974	0.815
5	0.98976	0.77378	0.96096	0.89824	0.98970	0.77383
10	0.99990	0.59874	0.99700	0.85414	0.99989	0.59891

True sensitivity and specificity of each component test are $S = 0.60$ and $C = 0.95$, respectively.

Copyright © 2015 John Wiley & Sons, Ltd.

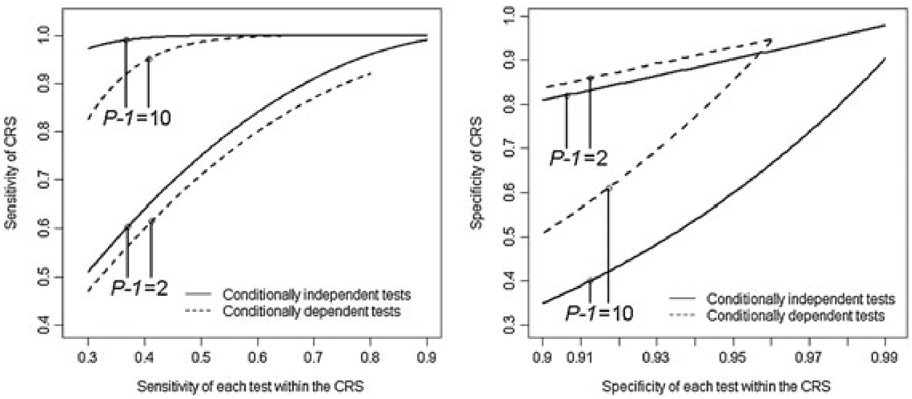


Figure 1. Sensitivity and specificity of the composite reference standards versus sensitivity and specificity, respectively, of component tests.

Figure 1 shows that for a given sensitivity of the component tests (x-axis), S_{CRS} increases as the number of component tests increases from $P - 1 = 2$ to $P - 1 = 10$. On the other hand, for a given specificity of the component tests, C_{CRS} decreases as $P - 1$ increases from 2 to 10.

4.2. Accuracy of composite reference standards under conditional dependence

To allow for comparison with the results under conditional independence, two particular settings are considered with greater and lesser conditional dependence. Conditional dependence within disease positive is created by setting $Pr(T_j = 1|L_1) \neq Pr(T_j = 1|L_2)$ and conditional dependence within disease negative by setting $Pr(T_j = 1|L_3) \neq Pr(T_j = 1|L_4)$. In setting (a), we consider a case where the magnitude of conditional dependence is greater. We set $Pr(T_j = 1|L_1) = 0.80, Pr(T_j = 1|L_2) = 0.40, Pr(T_j = 1|L_3) = 0.73$ and $Pr(T_j = 1|L_4) = 0.01$, $j = 1, \dots, P - 1$, so that the probability of a false negative outcome is substantially larger in L_2 than L_1 and the probability of a false positive outcome much larger in L_3 than L_4. In setting (b) $Pr(T_j = 1|L_1) = 0.61, Pr(T_j = 1|L_2) = 0.59, Pr(T_j = 1|L_3) = 0.06$ and $Pr(T_j = 1|L_4) = 0.0494$, corresponding to a situation where conditional dependence is weak. In both cases, the probabilities stated previously ensured $S = 0.60$ and $C = 0.95$.

From Table II and Figure 1, it can be seen that in setting (a) S_{CRS} does not increase as quickly when $P-1$ increases from 2 to 10, while C_{CRS} also does not decline as quickly compared with the case when the tests are conditionally independent. From Table II, under setting (b), changes in S_{CRS} and C_{CRS} resemble those observed under conditional independence. When the CRS is based on all conditionally dependent tests, it will never achieve a specificity of one. For example, under the conditional dependence setting (a), the component tests have a maximum possible specificity of 0.96, and C_{CRS} reaches a maximum value of 0.948 (Figure 1).

5. Bias in composite reference standards-based accuracy estimates

We now investigate the behavior of S_P^* and C_P^* in relation to the composition of the CRS. Of particular, interest is the impact of the sensitivity and specificity of the component tests and the true disease prevalence on the bias in CRS-based sensitivity and specificity ($S_P^* - S_P$, $C_P^* - C_P$). We consider the same ranges for accuracy of the component tests and true disease prevalence as in the previous section. The true values of the accuracy of the index test are set to $S_P = 0.90$ and $C_P = 0.90$. Under conditional dependence, these values are obtained by setting $Pr(T_P = 1|L_1) = 0.98$ and $Pr(T_P = 1|L_2) = 0.82$, resulting in $S_P = \frac{0.05\times0.98+0.05\times0.82}{0.05+0.05} = 0.9$. The probabilities contributing to the index test specificity are set to $Pr(T_P = 1|L_3) = 0.95$ and $Pr(T_P = 1|L_4) = 0.05$, resulting in $C_P = \frac{0.05\times0.05+0.85\times0.95}{0.05+0.85} = 0.9$.

5.1. Trends in CRS-based sensitivity and specificity with changes in accuracy of component tests and disease prevalence

Figure 2 presents CRS-based sensitivity (S_P^*) and specificity (C_P^*) for varying accuracy of the component tests, and disease prevalence assuming all tests are conditionally independent. The lines in each plot correspond to a different number of component tests ranging from 2 to 10. It should be noted that the values plotted are the expected values S_P^* and C_P^* (derived using Equations (5) and (6)).

Copyright © 2015 John Wiley & Sons, Ltd.

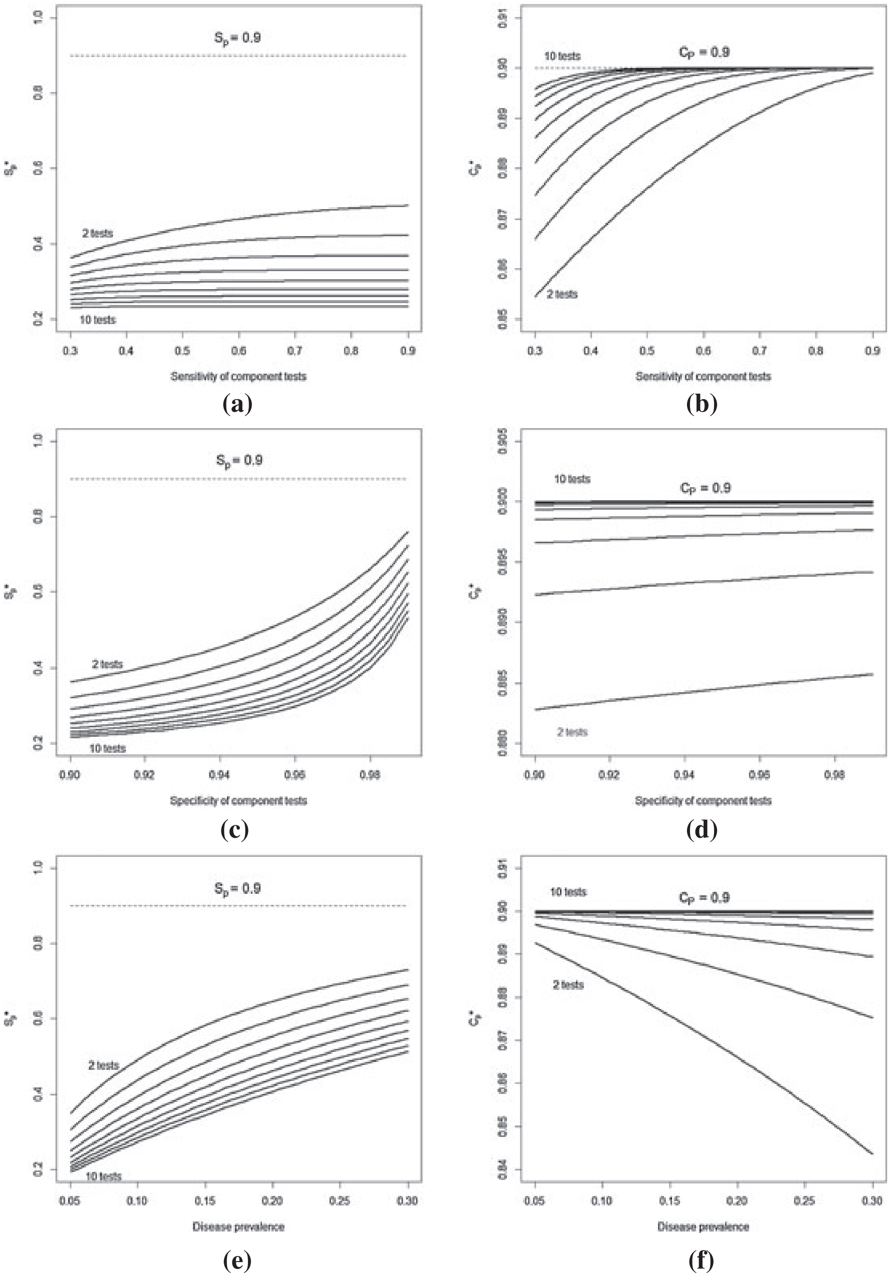


Figure 2. Composite reference standards-based sensitivity (S_P^*) and specificity (C_P^*) against accuracy of the component tests and disease prevalence while all tests are conditionally independent. Upper panel (a and b): change in $S = (0.30, 0.90)$, while $Pr(D = 1) = 0.10$, C = 0.95. Middle panel (c and d): change in $C = (0.90, 0.99)$, while $Pr(D = 1) = 0.10$, S = 0.60. Lower panel (e and f): change in $Pr(D = 1) = (0.05, 0.30)$, while C = 0.95, S = 0.60. Each curve corresponds to a composite reference standards with a different number of component tests $2 \leqslant (P - 1) \leqslant 10$).

From the three panels on the left of Figure 2, it can be seen that the CRS-based sensitivity S_P^* is much lower than the true value of $S_P = 0.90$ in all settings. We find that the bias in S_P^* is determined primarily by worsening specificity of the component tests and decreasing true disease prevalence (see sharper slopes in Figure 2(c) and e vs. Figure 2(a)). On the other hand, the bias in C_P^* is primarily due to lower sensitivity of the component tests and increasing prevalence (Figure 2(b) and f vs. Figure 2(d)). The bias illustrated in these figures corresponds to the particular case where the component tests share the same properties. From Equation (5), we can calculate precisely the bias from more general cases, such as when each successive component test in the CRS has better accuracy. For example, assume $Pr(D = 1) = 0.10$, and

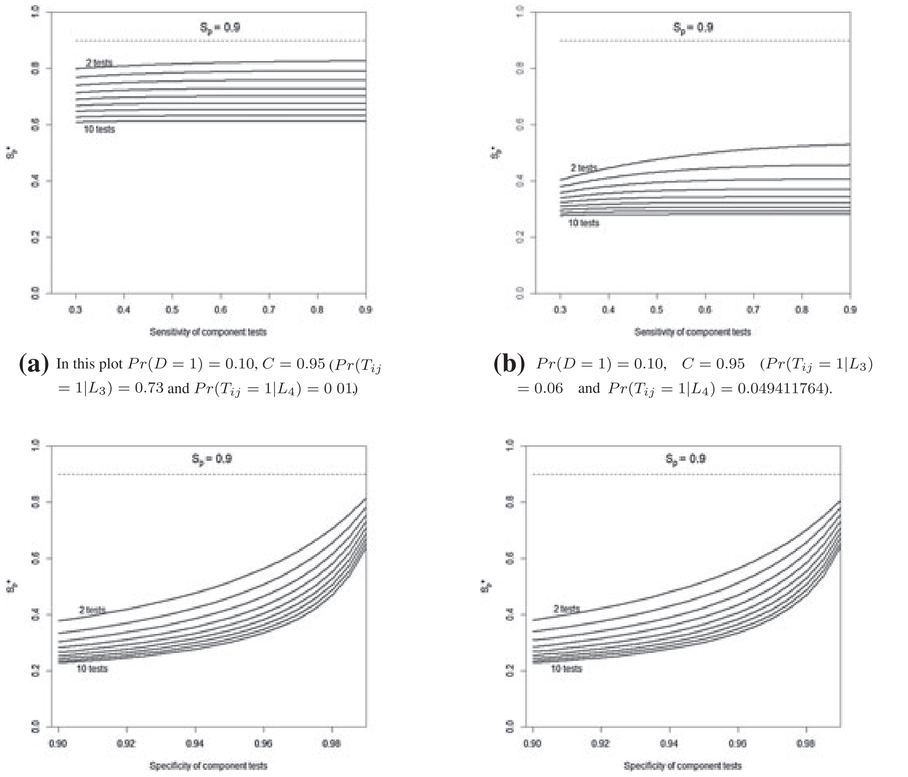
Copyright © 2015 John Wiley & Sons, Ltd.

that the second test in the CRS improves over the first one in terms of specificity such that $S_1 = S_2 = 0.6$, $C_1 = 0.94$ and $C_2 = 0.96$. From Equation (5), we can calculate that $S_3^* = 0.49$.

Figures 3 and 4 summarize the trends in S_P^* and C_P^*, respectively, when there is conditional dependence between the component tests of the CRS and the index test. The true disease prevalence is set at 0.10 in all scenarios. Greater conditional dependence within disease negative subjects tends to result in decreasing the amount of bias in S_P^* (Figure 3(a) vs. (b)). This can be explained by the observations of Table II where we found that the specificity of the CRS decreases more slowly with each additional test in the presence of conditional dependence. It should be noted that in other settings, it is also possible S_P^* will be overestimated [19]. For instance, changing only $Pr(T_j = 1|L_3) = 0.90$ and $Pr(T_j = 1|L_4) = 0$, it can be shown from Equation (11) that S_P^* will overestimated, tending to the limit of 0.916 with increasing number of component tests.

Figure 4 shows that greater conditional dependence within disease negative will also impact C_P^*, resulting in an overestimate compared with C_P (Figure 4(a) vs. (b)). The reason for the over-estimation can be understood by examining Equation (12). We set $Pr(T_j = 1|L_4) < Pr(T_j = 1|L_k)$ $k = 1, 2, 3$, implying that as the sensitivity of the component tests increases $Pr(T_j = 0|L_4) >> Pr(T_j = 0|L_k)$, $k = 1, 2$. Therefore, $P(T_{iP} = 0|CRS = 0)$ is over-estimated because of the increasing influence of $Pr(T_j = 0|L_4)$ with higher sensitivities. Similarly, greater conditional dependence within disease positive can also contribute to overestimation in C_P^* (Figure 4(c) and (d)).

For example, suppose that when the CRS comprises three conditionally independent tests having sensitivity $S = 0.90$ and specificity $C = 0.95$, and true disease prevalence is 0.10, $S_P^* = 0.45$ is considerably lower than the true value of S_P (Figure 2(a)) but $C_P^* = 0.8998 \approx C_P$ (Figure 2(b)). If the tests were conditionally dependent, the bias in S_P^* may be decreased (Figure 3); however, C_P^* will be an



(a) In this plot $Pr(D = 1) = 0.10, C = 0.95$ $(Pr(T_{ij} = 1|L_3) = 0.73$ and $Pr(T_{ij} = 1|L_4) = 0\ 01)$

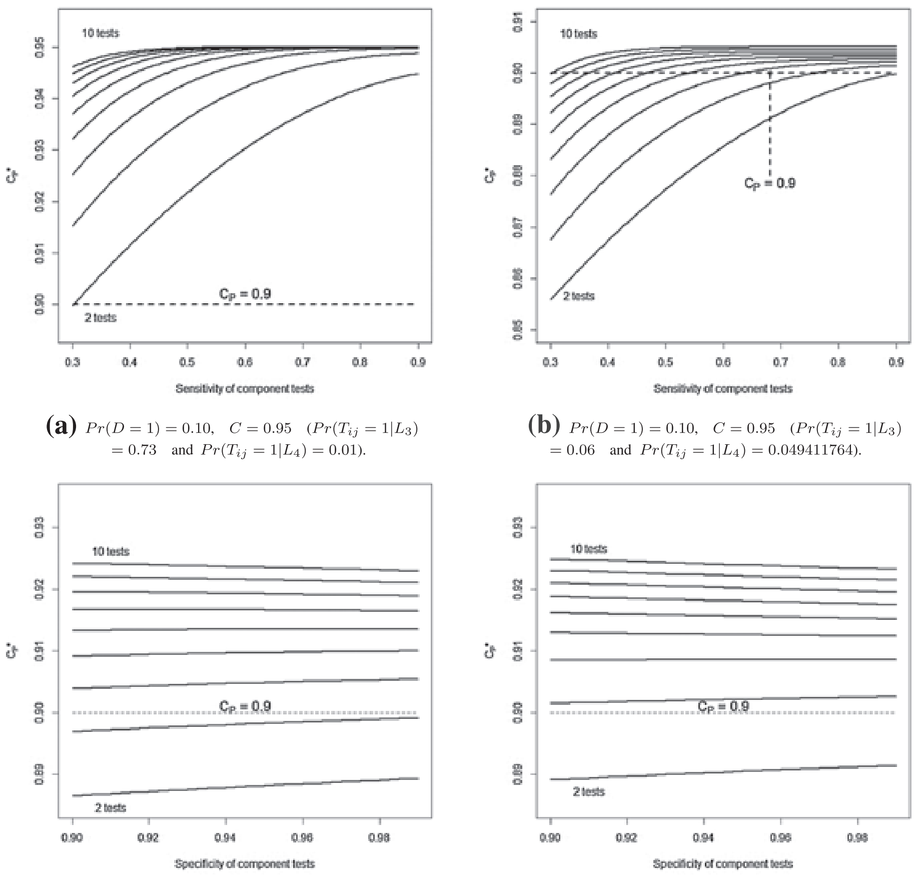
(b) $Pr(D = 1) = 0.10$, $C = 0.95$ $(Pr(T_{ij} = 1|L_3) = 0.06$ and $Pr(T_{ij} = 1|L_4) = 0.049411764)$.

(c) Change in $C = (0.90, 0.99)^‡$, while $Pr(D = 1) = 0.10$, $S = 0.60$ $(Pr(T_{ij} = 1|L_1) = 0.80$ and $Pr(T_{ij} = 1|L_2) = 0.40)$.

(d) Change in $C = (0.90, 0.99)^‡$, while $Pr(D = 1) = 0.10$, $S = 0.60$ $(Pr(T_{ij} = 1|L_1) = 0.61$ and $Pr(T_{ij} = 1|L_2) = 0.59)$.

Figure 3. Composite reference standards-based sensitivity (S_P^*) against sensitivity of component tests (upper panels) and specificity of compoonent tests (lower panels) while assuming conditional dependence between all tests. Each curve corresponds to a composite reference standards with a different number of component tests $2 \leqslant (P-1) \leqslant 10$. $Pr(T_j = 1|L_1) - Pr(T_j = 1|L_2) = 0.1$ in upper panel. $Pr(T_j = 1|L_3) - Pr(T_j = 1|L_4) = 0.053$ in lower panel.

Copyright © 2015 John Wiley & Sons, Ltd.



(a) $Pr(D = 1) = 0.10, \quad C = 0.95 \quad (Pr(T_{ij} = 1|L_3) = 0.73 \quad \text{and} \quad Pr(T_{ij} = 1|L_4) = 0.01).$

(b) $Pr(D = 1) = 0.10, \quad C = 0.95 \quad (Pr(T_{ij} = 1|L_3) = 0.06 \quad \text{and} \quad Pr(T_{ij} = 1|L_4) = 0.049411764).$

(c) $Pr(D = 1) = 0.10, \quad S = 0.60 \quad (Pr(T_{ij} = 1|L_1) = 0.80 \quad \text{and} \quad Pr(T_{ij} = 1|L_2) = 0.40).$

(d) $Pr(D = 1) = 0.10, \quad S = 0.60 \quad (Pr(T_{ij} = 1|L_1) = 0.61 \quad \text{and} \quad Pr(T_{ij} = 1|L_2) = 0.59).$

Figure 4. Composite reference standards-based specificity (C_P^*) against sensitivity of component tests (upper panels) and specificity of componente test (lower panels) while assuming conditional dependence between all tests. Each curve corresponds to a CRS with a different number of component tests $2 \leqslant (P-1) \leqslant 10$. $Pr(T_j = 1|L_1) - Pr(T_j = 1|L_2) = 0.1$ in upper panel. $Pr(T_j = 1|L_3) - Pr(T_j = 1|L_4) = 0.053$ in lower panel.

overestimate (Figure 4). Although the percentage bias in C_P^* is seemingly small, it could translate into a large underestimate of the false-positive percentage attributable to the index test in a low prevalence setting.

5.2. *Impact of increasing number of tests in the composite reference standards*

For the scenarios considered, as the number of component tests in the CRS increases, S_P^* tends to be increasingly underestimated (Figures 2 and 3). On the other hand, with increasing component tests, C_P^* may either be estimated better provided the tests are conditionally independent or tends to become overestimated if the tests are conditionally dependent (Figures 2 and 4).

As the number of component tests increases, more subjects will be classified by the CRS as having the disease. Although in practice the number of component tests $(P-1)$ in the CRS will be finite and relatively small, it is instructive to see what happens when the number of component tests increases infinitely $(P-1 \rightarrow \infty)$. Provided the set of conditions for achieving unbiased S_P^* and C_P^* listed under Equations 5 and 6 are not met, we will observe the asymptotic results described below.

From Equation (7), it can be seen that as $lim_{P-1\rightarrow\infty}\ Pr(CRS = 1) = 1$. Accordingly, under the assumption of conditional independence, it can also be shown that $lim_{P-1\rightarrow\infty}Pr(T_P = 1|CRS = 1) = Pr(T_P = 1) = 0.18$, which is substantially smaller than $S_P = 0.90$. Paradoxically, $lim_{P-1\rightarrow\infty}Pr(T_P = 0|CRS = 0) = Pr(T_P = 0|D = 0) = C_P$, meaning the CRS-based specificity of the index test converges to an unbiased estimate in a situation where in fact it is not possible to estimate specificity, as the number of subjects classified as disease negative converges to zero.

Copyright © 2015 John Wiley & Sons, Ltd.

A re-examination of Equation (5) helps understand the limiting behavior of S_P^*. It can be shown that $lim_{P-1\to\infty}\{1-\prod_{j=1}^{P-1}(1-Pr(T_j=1|D=d))\}=1$. Therefore, for a sufficiently large number of component tests, $S_P^*=Pr(T_P=1|CRS=1)\approx\sum_{d=0}^{1}Pr(T_P=1|D=d)Pr(D=d)=Pr(T_P=1)$, implying that as the number of component tests increases their accuracy has decreasing influence on the estimate of index test sensitivity. As we have chosen $C_j>1-S_j$, then $\prod_{j=1}^{P-1}Pr(T_j=0|D=0)=\prod_{j=1}^{P-1}C_j=C^{P-1}$ decreases toward zero much more slowly than the term $\prod_{j=1}^{P-1}Pr(T_j=0|D=1)=\prod_{j=1}^{P-1}(1-S_j)=(1-S_j)^{P-1}$. Therefore, all terms of Equation (6) except $Pr(T_P=0|D=0)$ become negligible and C_P^* tends toward the true specificity C_P as $P-1$ increases (even though it cannot be estimated).

When there is strong conditional dependence in the disease negative group, as the number of component tests in the CRS increases, $lim_{P-1\to\infty}\{\prod_{j=1}^{P-1}(1-Pr(T_j=1|L_k))\}=0,\ k=1,2,3,4$. However, the limit converges much faster for $k=1,2,3$ than for $k=4$, hence the index test specificity C_P^* converges to $Pr(T_P=0|CRS=0)\approx Pr(T_P=0|L_4)=0.95$, which is an over-estimate.

5.3. *Returning to evaluation of a test for C. trachomatis infection*

In the light of the preceding sections, we can see that the decline in $\hat{S}_{PCRU}^*$ and increase in $\hat{C}_{PCRU}^*$ seen in Table I is to be expected with increasing number of component tests in the CRS. Given CULC is supposed to have near perfect specificity but only moderate sensitivity, CULC positive patients are probably a subset of disease positive patients with a higher organism load. Because the sensitivity of the PCR test is also affected by the organism load, the sensitivities of the two tests are likely to be conditionally dependent. Thus, the CULC-based sensitivity estimate could be an over-estimate (if the conditional dependence was sufficiently high) or an underestimate. The CULC-based specificity estimate is probably an underestimate due to the imperfect sensitivity of CULC and may thus serve as a lower bound of the true PCR specificity.

The CRSs in Table I may be constructed in practice with the expectation they improve over CULC and will add insight, particularly by narrowing the range of uncertainty around the accuracy of the specificity of PCR. However, as we can see from our simulations, with each additional component test CRS, sensitivity will improve but with a loss of specificity, given the component tests in the example in Table I do not have perfect specificity. Therefore, we can expect that increasing $P-1$ will probably result in increasing underestimation of $\hat{S}_{PCRU}^*$. Further, with the addition of NAATs to the CRS, we can expect conditional dependence within disease negative of the CRS and the PCR test. Therefore, the $\hat{C}_{PCRU}^*$ values in Table I could be either underestimates or overestimates although it is unknown at what value of $P-1$ the direction of bias changes. Thus, although the definition of a CRS may be considered 'transparent', the resulting estimates are highly likely to be biased and are not easily interpretable even as lower or upper bounds of the new test's accuracy.

6. Other decision rules for defining a CRS

6.1. *Composite reference standards based on the AND decision rule*

While our focus so far has been on the OR rule motivated by the case of *C. trachomatis* testing, other alternative compositions of the CRS exist. For other applications, the OR-rule may be seen as being too liberal requiring only a single positive component test to classify a patient as true positive. At the other extreme, a CRS based on the AND rule (denoted CRS_a) would be very conservative requiring all tests to be positive as follows

$$CRS_a=I(T_1,\dots,T_{P-1})=\begin{cases}1 & \text{if } min(T_1,\dots,T_{P-1})=1\\ 0 & \text{if } min(T_1,\dots,T_{P-1})=0,\end{cases}$$

As we will show below, the expressions for the accuracy of this CRS and for the estimated accuracy of an index test with respect to this CRS are symmetric to the expressions defined previously in Section 3. By replacing C_P by S_P, S_P by C_P, C_j by S_j, S_j by C_j, and $1-Pr(D=1)$ by $Pr(D=1)$ in Equations (3), (4), (5), and (6), we can obtain the corresponding expressions for the case when all tests are conditionally independent. The sensitivity of CRS_a is given by

$$\begin{aligned} S_{CRS_a} &= Pr(CRS_a = 1|D = 1) = Pr(min(T_1, \dots, T_{P-1}) = 1|D = 1) \\ &= Pr(T_1 = 1, \dots, T_{P-1} = 1|D = 1) \\ &= \prod_{j=1}^{P-1} Pr\left(T_j = 1|D = 1\right) = \prod_{j=1}^{P-1} S_j, \end{aligned} \tag{13}$$

the specificity of CRS$_a$ is given by

$$\begin{aligned} C_{CRS_a} &= Pr(CRS_a = 0|D = 0) = Pr(min(T_1, \dots, T_{P-1}) = 0|D = 0) \\ &= 1 - Pr(min(T_1, \dots, T_{P-1}) = 1|D = 0) = 1 - Pr(T_1 = 1, \dots, T_{P-1} = 1|D = 1) \\ &= 1 - \prod_{j=1}^{P-1}(1 - Pr(T_j = 0|D = 0)) = 1 - \prod_{j=1}^{P-1}(1 - C_j), \end{aligned} \tag{14}$$

the sensitivity of index test P with respect to CRS$_a$ is given by

$$\begin{aligned} S^*_{Pa} &= Pr(T_P = 1|CRS_a = 1) \\ &= \frac{\sum_{d=0}^{1} Pr(T_P = 1|D = d)Pr(D = d)\prod_{j=1}^{P-1} Pr(T_j = 1|D = d)}{\sum_{d=0}^{1} Pr(D = d)\prod_{j=1}^{P-1} Pr(T_j = 1|D = d)}, \end{aligned} \tag{15}$$

and the specificity of index test P with respect to CRS$_a$ is given by

$$\begin{aligned} C^*_{Pa} &= Pr(T_P = 0|CRS_a = 0) \\ &= \frac{\sum_{d=0}^{1} Pr(T_P = 0|D = d)\left\{1 - \prod_{j=1}^{P-1}\left(1 - Pr\left(T_j = 0|D = d\right)\right)\right\} Pr(D = d)}{\sum_{d=0}^{1}\left\{1 - \prod_{j=1}^{P-1}\left(1 - Pr\left(T_j = 0|D = d\right)\right)\right\} Pr(D = d)}. \end{aligned} \tag{16}$$

Because of the symmetry, S_{CRS_a} decreases, and C_{CRS_a} increases as the number of component tests becomes large. If we were to use an AND decision rule for the situation in Table II (i.e. each component tests having $S = 0.6$ and $C = 0.9$), S_{CRS_a} would drop from 0.36 with 2 component tests to 0.0060 with 10 component tests. Correspondingly, C_{CRS_a} would increase from 0.9975 to 1. The expressions in Section 3.2 can similarly be converted into equivalent expressions for CRS_a.

6.2. Composite reference standards based on the ‘at least two positive tests’ decision rule

The OR and AND rules can be seen as the boundaries of a family of decision rules, which are defined by a positivity criterion of the form ‘at least m positive component tests’ as follows

$$CRS_m = I(T_1, \dots, T_{P-1}) = \begin{cases} 1 & \text{if } \sum_{j=1}^{P-1} T_j \geqslant m \\ 0 & \text{if } \sum_{j=1}^{P-1} T_j < m, \end{cases}$$

for $1 < m < P - 1$. We can see that if $m = 1$ we have the OR-decision rule and if $m = P - 1$ the AND-decision rule. We will examine in some detail the rule that requires at least two positive tests to classify a patient as positive, that is, when $m = 2$. CRS_2 is defined as follows

$$CRS_2 = I(T_1, \dots, T_{P-1}) = \begin{cases} 1 & \text{if } \sum_{j=1}^{P-1} T_j \geqslant 2 \\ 0 & \text{if } \sum_{j=1}^{P-1} T_j < 2, \end{cases}$$

Assuming conditional independence, the sensitivity of CRS$_2$ is given by

$$\begin{aligned} S_{CRS_2} &= Pr(CRS_2 = 1|D = 1) = Pr\left(\sum_{j=1}^{P-1} T_j \geqslant 2 \,\middle|\, D = 1\right) \\ &= 1 - \prod_{j=1}^{P-1}\left(1 - Pr\left(T_j = 1|D = 1\right)\right) - \sum_{l=1}^{P-1} Pr\left(T_l = 1|D = 1\right) \prod_{l \neq j}\left(1 - Pr\left(T_j = 1|D = 1\right)\right) \\ &= 1 - \prod_{j=1}^{P-1}\left(1 - S_j\right) - \sum_{l=1}^{P-1} S_l \prod_{l \neq j}\left(1 - S_j\right), \end{aligned} \tag{17}$$

and the specificity of the CRS$_2$ is given by

$$\begin{aligned} C_{CRS_2} &= Pr(CRS_2 = 0|D = 0) = Pr\left(\sum_{j=1}^{P-1} T_j < 2 \,\middle|\, D = 0\right) \\ &= \prod_{j=1}^{P-1} Pr\left(T_j = 0|D = 0\right) + \sum_{l=1}^{P-1}(1 - Pr(T_l = 0|D = 0)) \prod_{l \neq j} Pr(T_j = 0|D = 0) \\ &= \prod_{j=1}^{P-1} C_j + \sum_{l=1}^{P-1}(1 - C_l) \prod_{l \neq j} C_j. \end{aligned} \tag{18}$$

The sensitivity of index test P with respect to the CRS$_2$ is given by

$$\begin{aligned} S^*_{P2} &= Pr(T_P = 1|CRS_2 = 1) \\ &= \frac{S_P Pr(D = 1) S_{CRS_2} + (1 - C_P)(1 - Pr(D = 1))(1 - C_{CRS_2})}{Pr(D = 1) S_{CRS_2} + (1 - Pr(D = 1))(1 - C_{CRS_2})}, \end{aligned} \tag{19}$$

and the specificity of index test P with respect to the CRS$_2$ is given by

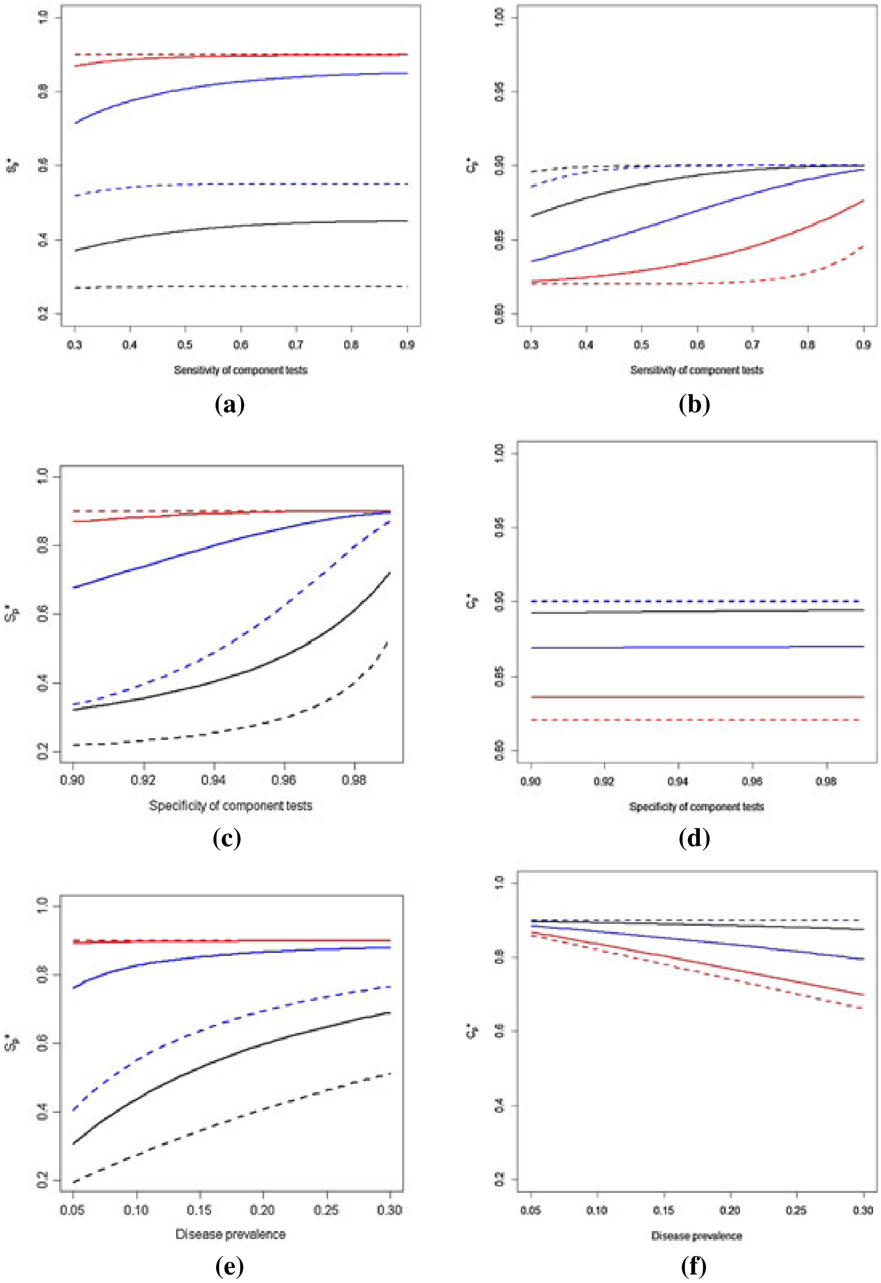
$$\begin{aligned} C^*_{P2} &= Pr(T_P = 0|CRS_2 = 0) \\ &= \frac{(1 - S_P) Pr(D = 1)(1 - S_{CRS_2}) + C_P(1 - Pr(D = 1)) C_{CRS_2}}{Pr(D = 1)(1 - S_{CRS_2}) + (1 - Pr(D = 1)) C_{CRS_2}}. \end{aligned} \tag{20}$$

As with the other decision rules defined previously, the sensitivity of CRS$_2$ depends only on the component tests' sensitivities, while the specificity of CRS$_2$ depends only on the component tests' specificities. Similar to the OR rule, as we increase the number of component tests, CRS_2's sensitivity will increase, while its specificity will decline, although the change is at a slower pace. For the settings in Table II, that is, each component test having $S = 0.60$ and $C = 0.95$, we have $S_{CRS2} = 0.36$ and $C_{CRS2} = 0.9975$. Note that this happens to be identical to the AND rule for the particular case of two component tests. If we increase the number of component tests to 10, then we would have $S_{CRS2} = 0.9983$ and $C_{CRS2} = 0.9139$, respectively.

6.3. Comparison of index test accuracy estimates based on different composite reference standards decision rules

Figure 5 compares the CRS-based estimates of sensitivity (S^*_P) and specificity (C^*_P) based on three different CRS decision rules: the OR rule (black lines), the AND rule (red lines), and the 'at least two positive tests' rule (blue lines). In each color, the solid and dashed lines correspond to a CRS based on 3 and 10 component tests, respectively. The true values of the index test accuracy are $S_P = 0.9$ and $C_P = 0.9$. In all plots, we see that the OR and AND rules define boundary estimates, and the 'at least two positive tests' rule is intermediate between them. The most bias in S^*_P (left panel) is observed under the OR decision rule (black curve), while the least bias can be seen under the AND rule (red curve), irrespective of the number of component tests used. By symmetry, the AND rule will create the most bias in C^*_P (right panel) while the OR decision rule will create the least bias, among the family of rules defined in Section 6.2.

Copyright © 2015 John Wiley & Sons, Ltd.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5. A comparison of estimated sensitivity (S_P^*) and estimated specificity (C_P^*) based on three different decision rules versus accuracy of the component tests and disease prevalence, while all tests are conditionally independent (true parameter values are $S_P = 0.9$ and $C_P = 0.9$ in all cases). Black lines = OR decision rule, blue line = at least two positive tests decision rule, red = AND decision rule. Solid line = CRS with three component tests, dashed line = CRS with 10 component tests. Upper panel (a and b): $Pr(D = 1) = 0.10$, $C = 0.95$. Middle panel (c and d): $Pr(D = 1) = 0.10$, $S = 0.60$. Lower panel (e and f): $C = 0.95$, $S = 0.60$.

7. Discussion

We studied in detail the performance of a CRS based on an OR decision rule. This type of CRS has been used in diagnostic research studies to improve the sensitivity in identifying the disease of interest over any single imperfect reference test. We showed that even if all component tests have excellent performance, for example, with 0.90 sensitivity and 0.95 specificity, the resulting estimates of index test accuracy may be highly biased. In practice, the magnitude and direction of the biases will be difficult to quantify precisely as they depend on the unknown accuracy of the component tests, the disease prevalence and the degree of conditional dependence between the tests.

Copyright © 2015 John Wiley & Sons, Ltd.

The definition of the CRS does not involve the test under evaluation. Therefore, it is perceived as being 'independent' of the test under evaluation [6, 8]. However, as we have shown, conditional dependence between component and index tests could arise due to their common dependence on a variable besides the true disease status. Intuitively, one can imagine that if the new test and the CRS systematically make the same errors (i.e., are conditionally dependent), the accuracy of the new test will be over-estimated. Further, we showed that CRS-based estimates of sensitivity and specificity are dependent on the true unknown disease prevalence. Therefore, the same CRS applied in different prevalence settings to evaluate the same index test would give different estimates of sensitivity and specificity, unlike the true sensitivity and specificity of the index test which are mathematically independent of disease prevalence.

Our simulation settings were inspired by those encountered in *C. trachomatis* testing, but the results may be generalized to other settings. It should be noted that other settings may result in different biases, for example, using the same sensitivities and specificities for the component tests as we have, but increasing the prevalence to > 0.50 could result in a greater magnitude of bias in S_P^* compared with C_P^*, and greater impact of conditional dependence within disease positive.

The sensitivity of a CRS based on the OR rule will increase with an increase in the number of component tests, although this will be at the cost of a loss of specificity unless all tests have perfect specificity. In the limiting case when the number of component tests tends to infinity, we found the paradoxical result that the estimated prevalence tends to 100%, while the CRS-based specificity tends to the true value of the specificity. Further, the estimated sensitivity tends to the probability of a positive test, suggesting the properties of the component tests in the CRS play no role in its estimation. Thus, as more information becomes available due to results of multiple tests being gathered, the performance of the CRS may worsen. This is in contrast to the expected performance of a well-defined statistical method. The apparent paradox can be explained by the fact that the CRS is overly simplistic in its construction. It ignores information on inter-relations between the component tests and reduces their joint results to a dichotomous result.

It has been argued that the CRS is clinically meaningful because it represents a clinical diagnosis and not the true disease status, which is impossible to determine in the absence of a gold-standard [20]. In a clinical setting, a physician faced with results of multiple imperfect tests may use them in a composite decision rule after weighing the risks of missed diagnosis versus overdiagnosis. But it is questionable whether the same composite rule should be applied to the evaluation of a new test or estimation of disease prevalence with no attempt to correct for the false-positive or false-negative errors in the CRS. It is rather like requiring that the new test replicate the errors of the CRS. Further, clinical diagnosis is more complex than a simple composite decision rule, taking into consideration additional variables, for example disease history, and the particular combination of positive and negative results.

Based on our findings, we recommend that a CRS based on combining results of NAATs via an OR or AND decision rule (or other variations) should not be used for estimating the accuracy of new *C. tracomatis* tests as follows: (i) the component tests are not guaranteed to have perfect sensitivity or specificity; (ii) this approach would ignore the conditional dependence between the tests; (iii) this approach would ignore the inter-relation between the different component tests and the index test.

The problem of evaluating a new test or estimating disease prevalence in the absence of a gold-standard reference remains a challenging one. CRSs have been promoted as a better approach than alternatives like latent class models [6, 20]. Yet, as we have shown, CRSs based on an any-positive (or all positive rule) suffer from a number problems that have not been acknowledged previously. Other CRSs based on more complex any-positive rules have also been found to result in bias [16]. We conclude that future research in this area should be directed towards approaches based on realistic statistical modeling of the observed data. Such models should take into account: (i) the inter-relations between all component tests and the index test; (ii) model conditional dependence between all tests (component and index); (iii) model disease prevalence; and (iv) incorporate external information if available, thus making complete use of the collected data while acknowledging all the different parameters (prevalence and individual test accuracies) that may come into play [29–31]. While there are acknowledged challenges with estimating more complex statistical models [31, 32], improving our understanding of them is necessary to make optimal use of the data gathered.

Acknowledgements

This work was supported by the Canadian Institutes of Health Research (Grant Number 89857). Nandini Dendukuri is supported by a Chercheur Boursier Senior award from the Fonds de Recherche en Santé - Québec. We gratefully acknowledged financial contribution by the Netherlands Organisation for Scientific Research (project 918.10.615).

Copyright © 2015 John Wiley & Sons, Ltd.

References

1. Reitsma JB, Rutjes AWS, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of Clinical Epidemiology* 2009; **62**(8):797–806.
2. Boyko EJ, Alderman BW, Baron AE. Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. *Journal of General Internal Medicine* 1988; **3**(5):476–481.
3. Whiting PF, Rutjes AWS, Westwood ME, Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology* 2013; **66**(10):1093–1104.
4. Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. *Journal of Chronic Diseases* 1981; **34**(12):599–610.
5. Black C. Current methods of laboratory diagnosis of Chlamydia trachomatis infections. *Clinical Microbiological Reviews* 1997; **10**:160–184.
6. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* 1999; **18**(22):2987–3003.
7. Johnson RE, Green TA, Schachter J, Jones RB, Hook EW, Black CM, Martin DH, Louis MES, Stamm WE. Evaluation of nucleic acid amplification tests as reference tests for chlamydia tratchomatis infections in asymptomatic men. *Journal of Clinical Microbiology* 2000; **38**(12):4382–4386.
8. Black CM, Marrazzo J, Johnson RE, Hook EW, Jones RB, Green TA, Schachter J, Stamm WE, Bolan G, St Louis ME, Martin DH. Head-to-head multicenter comparison of DNA probe and nucleic acid amplification tests for Chlamydia trachomatis infection in women performed with an improved reference standard. *Journal of Clinical Microbiology* 2002; **40**(10):3757–3763.
9. Centers for Disease Control and Prevention. Screening test to detect *chlamydia trachomatis* and *neisseria gonorrhoeae* infections. *Morbidity and Mortality Weekly Report* 2002; **51 (RR15)**:1–39.
10. Shrier L, Dean D, Klein E, Harter K, Rice P. Limitations of screening tests for the detection of Chlamydia trachomatis in asymptomatic adolescent and young adult women. *American Journal of Obstetrics and Gynecology* 2004; **190**(3):654–662.
11. Baughman A, Bisgard K, Cortese M, Thompson W, Sanden G, Strebel P. Utility of composite reference standards and latent class analysis in evaluating the clinical accuracy of diagnostic tests for pertussis. *Clinical and Vaccine Immunology: CVI* 2008; **15**(1):106–114.
12. Denkinger CM, Schumacher SG, Boehme CC, Dendukuri N, Pai M, Steingart KR. Xpert MTB/RIF assay for the diagnosis of extrapulmonary tuberculosis: a systematic review and meta-analysis. *European Respiratory Journal* 2014; **44**:435–446.
13. Sinclair A, Xie X, Teltscher M, Dendukuri N. Systematic review and meta-analysis of a urine-based pneumococcal antigen test for diagnosis of community-acquired pneumonia caused by Streptococcus pneumoniae. *Journal of Clinical Microbiology* 2013; **51**:2303–2310.
14. Lind-Brandberg L, Welinder-Olsson C, Lagergård T, Taranger J, Trollfors B, Zackrisson G. Evaluation of PCR for diagnosis of Bordetella pertussis and Bordetella parapertussis infections. *Journal of Clinical Microbiology* 1998; **36**:679–683.
15. Naaktgeboren C, Bertens L, van Smeden M V, de Groot J, Moons K, Reitsma J. Value of composite reference standards in diagnostic research. *British Medical Journal* 2013; **347**:f5605.
16. Hadgu A, Dendukuri N, Wang L. Evaluation of screening tests for detecting Chlamydia trachomatis: bias associated with the patient-infected-status algorithm. *Epidemiology (Cambridge, Mass.)* 2012; **23**(1):72–82.
17. Black M, Craig B. Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* 2002; **21**(18):2653–69.
18. US Food and Drug Administration. Draft guidance for industry and food and drug administration staff – establishing the performance characteristics of in vitro diagnostic devices for Chlamydia trachomatis and/or Neisseria gonorrhoea: Screening and Diagnostic Testing, 2011. Available from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm254813.htm [Accessed on 6 November 2015].
19. Dendukuri N, Wang L, Hadgu A. Evaluating diagnostic tests for chlamydia trachomatis in the absence of a gold standard: a comparison of three statistical methods. *Statistics in Biopharmaceutical Research* 2011; **3**(2):385–397.
20. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
21. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2007; **8**(1465–4644 (Print)):474–484.
22. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* 1985; **41**(4):959–968.
23. Torrance-Rynard V, Walter S. Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* 1997; **16**:2157–2175.
24. Brenner H. How independent are multiple ‘independent’ diagnostic classifications? *Statistics in Medicine* 1996; **15**(0277–6715):1377–1386.
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. *Statistics in Medicine* 2012; **31**(11–12):1129–1138.
26. Centers for Disease Control and Prevention. Recommendations for laboratory-based detection of *chlamydia trachomatis* and *neisseria gonorrhoeae* - 2014. *Morbidity and Mortality Weekly Report* 2014; **63**((RR02)):1–19. (Available from http://www.cdc.gov/mmwr/preview/mmwrhtml/rr6302a1.htm [Accessed on 11 June 2015]).
27. Persing D. *Molecular Microbiology*. ASM Press: Washington, DC, 2004.
28. Cook RL, Hutchison SL, Østergaard L, Braithwaite RS, Ness RB. Systematic review: noninvasive testing for Chlamydia trachomatis and Neisseria gonorrhoeae. *Annals of Internal Medicine* 2005; **142**(11):914–25.
29. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* 1995; **141**(3):263–273.
30. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001; **57**(1):158–167.

Copyright © 2015 John Wiley & Sons, Ltd.

31. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Statistics in Medicine* 2009; **28**(0277-6715 (Print)):441–461.
32. van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KGM, de Groot JAH. Latent class models in diagnostic studies when there is no reference standard-a systematic review. *American Journal of Epidemiology* 2014; **179**(4):423–431.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

Copyright © 2015 John Wiley & Sons, Ltd.