Published in final edited form as:

Eur J Clin Microbiol Infect Dis. 2012 September ; 31(9): 2111–2116. doi:10.1007/s10096-012-1602-1.

Methods and recommendations for evaluating and reporting a new diagnostic test

A. S. Hess, M. Shardell, J. K. Johnson, K. A. Thom, P. Strassle, G. Netzer, and A. D. Harris
University of Maryland School of Medicine, Baltimore, MD, USA

A. S. Hess: ahess@epi.umaryland.edu

Abstract

No standardized guidelines exist for the biostatistical methods appropriate for studies evaluating diagnostic tests. Publication recommendations such as the STARD statement provide guidance for the analysis of data, but biostatistical advice is minimal and application is inconsistent. This article aims to provide a self-contained, accessible resource on the biostatistical aspects of study design and reporting for investigators. For all dichotomous diagnostic tests, estimates of sensitivity and specificity should be reported with confidence intervals. Power calculations are strongly recommended to ensure that investigators achieve desired levels of precision. In the absence of a gold standard reference test, the composite reference standard method is recommended for improving estimates of the sensitivity and specificity of the test under evaluation.

Introduction

With the rapid expansion of molecular diagnostics in infectious disease, clinical microbiologists are often called upon to review the literature for new diagnostic tests or to validate new tests in their own laboratory. Publication recommendations such as the STAndards for the Reporting of Diagnostic accuracy studies (STARD) provide excellent guidance on the design and report of a study, but biostatistical advice is minimal and the application of these guidelines is inconsistent in the literature [1]. At present, no standardized guidelines exist on biostatistical aspects of study design and reporting, and the investigator evaluating the new test is left to answer a number of questions. Should the sensitivity and specificity always be reported? What statistical tests or evaluations should be done? How many samples need to be assayed in order to generate useful and meaningful results? What reference test should be used, and what can be done if that test is not perfect? In the absence of accessible answers to some of these questions, many studies reporting new diagnostic tests miss opportunities to clarify and even expand the scope of their results.

Our objective is to provide a self-contained, accessible resource for researchers on the biostatistical aspects of evaluating and reporting the performance of dichotomous tests, i.e. tests that classify samples into two categories. In order to meet this goal, we will review and make recommendations for reporting on: (1) statistical methods for estimating sensitivity and specificity and their corresponding confidence intervals, (2) criteria for determining sample size in the evaluation of a new diagnostic test, and (3) statistical methods for evaluating a test when no gold standard is available. These recommendations do not stand

alone as guidelines for the reporting of diagnostic tests, but instead cover additional, important topics not addressed by existing common standards such as STARD.

The new test and the existing test

Comparisons of dichotomous tests are usually represented in a 2×2 “paired contingency” table (Fig. 1), with the new test in rows and the existing test in columns. If the existing test is assumed to be perfect it is called the ‘gold standard’. The values a , b , c , and d in Fig. 1a represent the number of samples corresponding to each of four possible pairs of outcomes. In the case of an existing gold standard, $(a + c)$ is the number of true positive samples, and $(b + d)$ is the number of true negative samples.

Sensitivity and specificity

Traditionally, the new test is compared to the existing test by the proportion of the true positive (‘sensitivity’) and true negative (‘specificity’) samples that it identifies. In Fig. 1a, the estimated sensitivity of the new test is $a/(a + c)$, and the estimated specificity is $d/(b + d)$.

The sensitivity and specificity of a test are not always fixed. Many dichotomous tests are based on assigning cutoff levels to a continuous scale, and altering those cutoff levels will alter the measured performance of the test. Generally, there is a trade-off between sensitivity and specificity as a cutoff level changes; for example, as the cutoff level changes and the sensitivity of a test increases, the specificity decreases. In addition, both sensitivity and specificity may be different under clinical conditions (‘diagnostic’ sensitivity and specificity) compared to ideal laboratory conditions (‘analytical’ sensitivity and specificity) [2]. Occasionally, a new test may be developed that is so obviously more sensitive or specific than any other existing test that it is inappropriate to compare them. It is not within the scope of this article to address the choice of samples, cutoff levels, or the circumstances in which a test should be applied. The biostatistical methods presented are valid for both diagnostic or analytical sensitivity and specificity, but the investigator should be wary of comparing tests performed under different conditions.

Example

Suppose a new test called “test A” has been developed to detect a pathogen in a clinical specimen. An investigator assays 100 samples in the laboratory using both the new test and a gold standard (Fig. 1b). According to the gold standard, 59 samples were truly pathogen-positive, and the remaining 41 samples were truly pathogen-negative. Test A correctly identified 54 of the 59 positive samples, so the estimated sensitivity of test A is $54/59 \times 100\% = 91.5\%$. Thirty-nine of the 41 negative samples were correctly identified, so the estimated specificity of test A is $39/41 \times 100\% = 95.1\%$.

Hypothesis tests such as McNemar’s or Liddell’s are sometimes used to compare a new test to an existing test, but they should never be used to assign a p-value to an estimate of sensitivity or specificity. McNemar’s and Liddell’s tests only examine the extent of disagreement between the two tests, i.e. they only use the values b , c , and N in their formulas. As shown above, the estimates of sensitivity and specificity also use a and d , the cells that represent agreement. Since McNemar’s and Liddell’s do not use all the information contained in the estimates of sensitivity and specificity, some other method for evaluation is needed.

Quantifying precision of sensitivity and specificity: confidence intervals

The precision of any estimates of sensitivity and specificity should be reported, no matter the value, in order to avoid misleading the reader about the results. For example, a coin flip

can be used as a diagnostic test with a ‘sensitivity’ of 50 %. Suppose an investigator flips a coin ten times, and the expected result is five heads. However, the probability of five heads in ten flips is only 0.25 [3]. Even if the investigator decides that a result of four or six heads is acceptable, the probability of getting any of those results in ten flips is only 0.66. In other words, the probability is 0.34 that the result of the ten-flip experiment will be a sensitivity that differs from the truth by 20 % or more. However if the coin is flipped 100 times, the chance of a result more than 20 % off falls to less than 0.0001. An experiment with one hundred flips is less likely to vary widely in its result, and so its estimate of the sensitivity is more precise.

A 95 % confidence interval can summarize the precision of an estimate by placing it in a range of values consistent with the data. The more narrow the confidence interval, the more precise the estimate. Figure 2 presents two imaginary tests that are both estimated to have a sensitivity of 60 %. The 95 % confidence interval for test one is (52 %, 68 %) and the interval for test two is (40 %, 80 %). The 95 % confidence interval for sensitivity of test one is narrower than that for test two, therefore the estimate is more precise. The formula for calculating 95 % confidence intervals for proportions like sensitivity and specificity is presented in Fig. 3 [3]. 95 % confidence intervals can be used to quantify precision in a way that is repeatable and widely understood, and the FDA has recommended that all estimates of sensitivity and specificity be reported with 95 % confidence intervals [4].

The width of a confidence interval relative to the whole possible range is often summarized by calculating the ‘margin of error’ M . The margin of error is simply one-half the width of the confidence interval, i.e. $(upper\ limit - lower\ limit)/2$. There is no commonly accepted standard for how small the margin of error should be. A margin of five percentage points is frequently used in the social sciences, but no such guideline exists in the medical literature [5–7]. In the absence of official recommendations from regulating agencies, we can only say that smaller margins of error are preferred.

Example

The estimated sensitivity of the new assay is 91.5 %, based on the results of 59 samples in Fig. 1b. The 95 % confidence interval for the sensitivity is (84.4 %, 98.6 %). The estimated specificity of the assay is 95.1 %, and the confidence interval for the specificity is (89.6 %, 100 %). The margin of error M for the sensitivity is $(0.986 - 0.844)/2 = 0.071$. The margin of error M for the specificity is $(1.006 - 0.896)/2 = 0.055$. These confidence intervals represent a range of plausible values for the sensitivity and specificity of test A. Note that the upper limit of the confidence interval for the specificity is greater than 100 %: this is an artifact of the confidence interval formula and does not mean that the specificity may be better than the specificity of the reference test.

Designing experiments to achieve precise estimates of sensitivity and specificity

In the previous example, the estimates of sensitivity and specificity were not very precise: both covered around 20 % of the entire range of possible values from 0 to 100 %. An investigator designing a study may want a certain level of precision in order for the results to be microbiologically and clinically meaningful. In order to achieve some pre-specified margin of error, a minimum number of samples will need to be tested. Separate sample sizes must be calculated for estimating the sensitivity and specificity, since only true-positive samples will be useful for estimating the sensitivity and only true-negative samples will be useful for estimating the specificity. It is also important that the individual samples are independent, i.e. they should not include multiple samples from the same subject or repeated

testing of the same sample. Small pilot studies are useful for obtaining preliminary estimates of sensitivity and specificity for use in sample size calculations.

Sample size calculations for diagnostic tests need to account for (1) the desired margin of error M , (2) the estimated (or presumed) sensitivity or specificity, and (3) the likelihood that the next study will find a different estimate of the sensitivity or specificity by chance. The previous example estimated that the sensitivity of test A was 91.5 %, but the 95 % confidence interval suggests that a sensitivity of 84.4 % is also plausible. Including this potential for variability in sample size calculations is analogous to including ‘power’ when determining sample sizes for randomized clinical trials. For diagnostic tests, the aim is to calculate a sample size that will have a particular probability (power) of estimating the sensitivity or specificity with a margin of error no larger than M . We recommend that sample size calculations have 90 % power. Exact power calculations for dichotomous tests are complicated, but a three-step method is presented below that gives a good approximation and only requires a calculator [8]. Corresponding equations are presented in Fig. 4.

Step 1: Calculate a rough estimate of the sample size n using a standard sample size formula [3]. Step 2: Calculate a 90 % confidence interval for sensitivity or specificity, and choose the bound closer to 0.5 (usually the lower). Call this boundary p^* . This is the step that specifies the power. Step 3: Repeat step 1 using p^* . The value n^* is the approximate number of samples needed for 90 % power to achieve a confidence interval with a margin of error M given a certain sensitivity or specificity.

Example

An investigator wants to calculate the sample size necessary for 90 % power to estimate a sensitivity of 91.5 % for test A with a relative margin of error 0.05. Following the steps presented previously, the investigator calculates (1) a preliminary sample size of 119.2, (2) a 90 % lower bound p^* of 86.9 %, and (3) a final sample size of 156 true-positive samples. A similar calculation for the specificity using the original estimate of 95.1 % yields a final sample size of 117 true-negative samples. The reader may notice that in this example the result for step 1 seems close enough to the final answer in step 3 that the remaining steps are not necessary—this is true for a sensitivity or specificity closer to 50 %, but as values get very close to 100 % the first step becomes an increasingly grave underestimate and the second and third steps become necessary. For example, if sensitivity were 97 %, then $n=45$ in step 1, but $n^*=91$ in step 3.

Evaluating a new diagnostic test in the absence of a gold standard

A true gold standard test is rarely available. For example, nucleic acid amplification tests from vaginal swabs are currently the best diagnostic test for *Chlamydia trachomatis*, but since the specificity is only 93 % this method is not a true gold standard [9]. Any time an imperfect test is used as a reference, disease status may be misclassified, resulting in an unpredictably biased estimate of the sensitivity and specificity [10, 11]. A variety of methods have been proposed to compensate for an imperfect reference test, all of which require one or more tests *in addition* to the new test and the imperfect reference standard. The additional test is required to provide new information about the samples that can be used to improve the estimates of sensitivity and specificity. If no additional tests are available to provide new information, there is no statistical method that will improve the estimates of sensitivity and specificity. We recommend the composite reference standard (CRS) proposed by Miller and Hadgu because, as we will demonstrate below, it is relatively easy to use and avoids some of the biases and assumptions of other methods [10, 12, 13].

The composite reference standard method uses three separate tests. The first test is the new test under investigation (N). The second test is an imperfect 'reference' test (S). This test is not treated as a gold standard but should have the best available sensitivity and specificity. Both tests are applied to all the samples in the pool. CRS then applies a third test to all the samples that tested negative under the reference test, i.e. the right-hand column of the 2×2 table. This test is called the 'resolver' (R). Even if the underlying method of the test is different from the reference test (e.g. a PCR based on a different gene) the resolver should have a sensitivity and specificity similar to the reference test. Consider the general case and the example presented in Fig. 5. The samples that tested negative by reference (b and d) are re-tested with the resolver. All samples that then test positive (a' and c') by the resolver are moved into the positive column, and all originally negative samples that test negative (equal to $b-a'$ and $d-c'$) retain their original classification (Fig. 5a).

The adjusted sensitivity and specificity of the new test according to the CRS method are presented in Fig. 6. To understand these equations, examine Fig. 5 and note that samples that test positive using the imperfect reference test *or* the imperfect resolver test ($a + c + a' + c'$) are treated as if they are true positives in the sensitivity calculation. Samples that test negative using the imperfect reference test *and* the imperfect resolver test ($b + d - a' - c'$) are treated as if they are true negatives in the specificity calculation. 95 % confidence intervals for CRS estimates of sensitivity and specificity can be calculated using the formulas presented in Fig. 6b [13, 14].

The CRS method assumes that the imperfect reference (S) is less sensitive than it is specific, i.e. it is more likely to classify a truly positive sample as negative than to classify a negative sample as positive. This is the case with many clinical tests. If there is reason to believe that the imperfect reference is less specific than it is sensitive, it may be appropriate to apply the resolver to the left-hand column of the 2×2 table, and reclassify the results accordingly.

Example

Test A is being compared to an imperfect gold standard (S) using the enlarged sample sizes estimated in the previous example. A third test (R) will be used as a resolver according to the CRS method. Using data in Fig. 5b and formulas in Fig. 6a, the CRS estimates of the sensitivity and specificity of test A are 91.4 % and 99.1 %. Note that the estimates of the sensitivity and specificity using only tests N and S are 91.7 % and 94.9 %, respectively. The CRS method suggests that the true sensitivity of test A may not be better than original estimate, but that the true specificity is much closer to 100 %. The 95 % confidence interval for sensitivity using CRS is (87.0 %, 95.7 %); for the CRS specificity it is (95.7 %, 100.6 %).

Other methods for evaluating a new diagnostic test in the absence of a gold standard

Other methods for improving the sensitivity and specificity exist in the microbiological literature. One of the most common is discrepant analysis, which is similar to CRS but applies one or more resolver tests to the cells in which the new test and the standard disagree (b and c in Fig. 1) [13]. A fundamental problem with discrepant analysis is that the 'improved' reference depends on the results of the new test, e.g., the samples in cell b are only distinguishable from the samples in cell d because they are positive by the new test. Evaluating a test by using its own results as a reference is unsatisfactory. Although discrepant analysis is intuitively appealing, it produces estimates with biases of unpredictable magnitude and direction [10–13, 15]. By contrast, CRS only applies the resolver test to samples that tested negative under the *reference* test, and uses no information

from the new test. The FDA strongly discourages any investigator from using discrepant analysis to estimate the sensitivity and specificity of a test [4].

Latent class analysis is another method for estimating the sensitivity and specificity of a new test in the absence of a gold standard. Latent class analysis uses a statistical model to estimate the true status of each sample when at least three imperfect tests have been used. Alonzo and Pepe have noted that latent class analysis has three drawbacks in the clinical setting. First, the presence of the disease is not explicitly defined by the latent class algorithm, but is instead an unmeasured or 'latent' variable. Second, it assumes that the test results are statistically independent given the 'true' infection status, and this assumption cannot be tested. Third, the estimates of sensitivity and specificity produced by the model are not clear arithmetic functions of the data [13, 16, 17]. The second limitation is of particular concern since test results from specimens are frequently positively correlated, and this tends to produce overestimates of sensitivity and specificity. Practically, latent class analysis also has the additional drawback that it requires every sample to be tested by at least three methods. The workings of latent class analysis are beyond the scope of this article, but we provide references for interested readers [17–20].

Conclusion and recommendations for reporting

Whenever a new dichotomous diagnostic is evaluated, the investigators should endeavor to compare it to an existing test and the sensitivity and specificity of the new test compared to the old should be reported. Including confidence intervals with these estimates allows both the accuracy of the test and the precision of the estimates to be evaluated. Although no standard for the width of confidence intervals exists in the biostatistical literature, narrower intervals are better. Readers and investigators should utilize *a priori* knowledge to determine reasonable confidence intervals to interpret results.

Prior to embarking on a study, the investigator should design the study using sample size calculations to ensure that sensitivity and specificity are precisely estimated, i.e. the confidence intervals are narrow. Using the formulas presented here, the desired precision and power can be specified ahead of time and the necessary number of samples calculated. As the examples presented demonstrate, if high precision and power are desired, then a considerable number of samples are necessary. Without power calculations, the investigator may be left with imprecise estimates through small sample numbers.

While it is convenient to assume that the reference test used in calculating these estimates is a perfect gold standard, this is rarely the case in practice. Several methods exist to reduce the bias that results from estimating the sensitivity and specificity of a new test using an imperfect reference, all of which require that a third test is applied to some or all of the samples. Although none of these techniques are perfect, we recommend using the composite reference standard method both for its statistical properties and its relative ease of use.

In clinical research there is almost always a trade-off between statistical ideals and the practical realities of sample collection. We have attempted to provide a resource for the clinical microbiologist that will help clarify the results of any evaluation of a new test, and provide tools to improve future investigations.

Acknowledgments

MS's work on this article was supported by NIH grant 1K25AG034216.

ADH's work on this article was supported by NIH grant 1K24AI079040-01A1.

References

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med*. 2003; 138(1):40–44. [PubMed: 12513043]
2. Pfeifer, J., editor. *Molecular genetic testing in surgical pathology*. Lippincott Williams & Wilkins; Philadelphia: 2006.
3. Rosner, BA. *Fundamentals of biostatistics*. 6. Thomson Brooks Cole; Belmont, CA: 2006.
4. FDA. Statistical guidance on reporting results from studies evaluating diagnostic tests. 2011. Available from: <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm>. Updated 6 January 2011; cited 8 December 2011
5. Royse, D.; Thyer, BA.; Padgett, DK. *Program evaluation: An introduction*. 5. Wadsworth, Cengage Learning; Belmont, CA: 2010.
6. Royse, D. *Research methods in social work*. 5. Thomson Brooks Cole; Belmont, CA: 2008.
7. Sullivan, LM. *Essentials of biostatistics in public health*. 1. Jones and Bartlett; Sudbury, MA: 2008.
8. Price RM, Bonett DG. Confidence intervals for a ratio of two independent binomial proportions. *Stat Med*. 2008; 27(26):5497–508. [PubMed: 18781560]
9. Schachter J, McCormack WM, Chernesky MA, Martin DH, Van Der Pol B, Rice PA, et al. Vaginal swabs are appropriate specimens for diagnosis of genital tract infection with chlamydia trachomatis. *J Clin Microbiol*. 2003; 41(8):3784–3789. [PubMed: 12904390]
10. Miller WC. Bias in discrepant analysis: when two wrongs don't make a right. *J Clin Epidemiol*. 1998; 51(3):219–231. [PubMed: 9495687]
11. Hawkins DM, Garrett JA, Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. *Stat Med*. 2001; 20(13):1987–2001. [PubMed: 11427955]
12. Hadgu A. The discrepancy in discrepant analysis. *Lancet*. 1996; 348(9027):592–593. [PubMed: 8774575]
13. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med*. 1999; 18 (22):2987–3003. [PubMed: 10544302]
14. Baughman AL, Bisgard KM, Cortese MM, Thompson WW, Sanden GN, Strebel PM. Utility of composite reference standards and latent class analysis in evaluating the clinical accuracy of diagnostic tests for pertussis. *Clin Vaccine Immunol*. 2008; 15(1):106–114. [PubMed: 17989336]
15. Lipman HB, Astles JR. Quantifying the bias associated with use of discrepant analysis. *Clin Chem*. 1998; 44(1):108–115. [PubMed: 9550567]
16. Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. *Stat Med*. 1997; 16(19):2157–2175. [PubMed: 9330426]
17. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics*. 2007; 8(2):474–484. [PubMed: 17085745]
18. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Stat Med*. 1986; 5(1):21–27. [PubMed: 3961312]
19. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*. 1996; 52(3):797–810. [PubMed: 8805757]
20. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Stat Methods Med Res*. 1998; 7(4):354–370. [PubMed: 9871952]

A

		Existing Test		
		Positive	Negative	
New Test	Positive	a	b	$a + b$
	Negative	c	d	$c + d$
		$a + c$	$b + d$	N

B

		Gold Standard		
		Positive	Negative	
Test A	Positive	54	2	56
	Negative	5	39	44
		59	41	100

Fig. 1.
a A 2×2 paired contingency table for comparing the results of two tests on the same samples. **b** Results of a comparison between a new assay (“*Test A*”) and a gold standard assay

Fig. 2. 95 % confidence intervals around estimates of sensitivity. Both tests have a sensitivity of 60 %. Test one (*upper*) has a 95 % confidence interval of (52 %, 68 %). Test two (*lower*) has a 95 % confidence interval of (40 %, 80 %). The 95 % confidence interval around the estimate of the sensitivity for test one is narrower than for test two, therefore the estimate is more precise

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Fig. 3.

Formula for 95 % confidence intervals for sensitivity or specificity. $p_{\bullet}$ is the estimate of sensitivity or specificity, n is either the number of true-positive samples (for sensitivity) or the number of true-negative samples (for specificity). This formula is appropriate as long as both $np_{\bullet}$ and $n(1-p_{\bullet})$ are not less than 5

Step 1:
$$n = \left(\frac{1.96}{M} \right)^2 \cdot \hat{p}(1 - \hat{p})$$

Step 2:
$$p^* = \hat{p} - 1.282 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Step 3:
$$n^* = \left(\frac{1.96}{M} \right)^2 \cdot p^*(1 - p^*)$$

Fig. 4.

Three-step method to approximate the sample size n^* with 90 % power to estimate p with a margin of error no more than M . Step 1 calculates a preliminary estimate n based on $p_{\bullet}$, the estimated sensitivity or specificity and M . Step 2 gives ‘power’ to the sample size estimate by calculating p^* , or the 90 % lower bound around $p_{\bullet}$ given n . Step 3 calculates n^* using the same equation as step 1, but substituting p^* for $p_{\bullet}$.

A

**New Test
(N)**

Imperfect Reference (S)				Imperfect Resolver (R)						
		S +	S -			S +	S -, R +	R -, S -		
	N +	a	b	a + b	→	a + a'	b - a'	a + b		
	N -	c	d	c + d		c + c'	d - c'	c + d		
		a + c	b + d	N			(a + c) + (a' + c')	(b + d) - (a' + c')		

B

**Test A
(N)**

Imperfect Gold Standard (S)				Imperfect Resolver (R)						
		S +	S -			S +	S -, R +	R -, S -		
	N +	143	6	149	→	143 + 5 (=148)	6 - 5 (=1)	149		
	N -	13	111	124		13 + 1 (=14)	111 - 1 (=110)	124		
		156	117	273			162	111		

Fig. 5.

a Summary of the two stages of a composite reference standard (CRS) test of a new test (N). Samples labeled negative by the imperfect standard (S) are re-tested with the third test, the imperfect 'resolver' (R). **b** Example showing the two stages of a CRS resolution of the new test, "Test A"

$$\text{CRS Sensitivity} = \hat{p}_{sens} = \frac{a + a'}{(a + c) + (a' + c')}$$

$$\text{CRS Specificity} = \hat{p}_{spec} = \frac{d - c'}{(b + d) - (a' + c')}$$

$$95\% \text{ CI for CRS Sensitivity} = \hat{p}_{sens} \pm 1.96 \sqrt{\frac{\hat{p}_{sens}(1 - \hat{p}_{sens})}{(a + c) + (a' + c')}}}$$

$$95\% \text{ CI for CRS Specificity} = \hat{p}_{spec} \pm 1.96 \sqrt{\frac{\hat{p}_{spec}(1 - \hat{p}_{spec})}{(b + d) - (a' + c')}}}$$

Fig. 6.

a Formulas for calculating sensitivity and specificity using a composite reference standard method. **b** Formulas for calculating 95 % confidence intervals around composite reference standard (CRS) estimates of sensitivity and specificity. See Fig. 5a for reference