Concerns about composite reference standards in diagnostic research

Nandini Dendukuri,[1] Ian Schiller,[1] Joris de Groot,[2] Michael Libman,[3] Karel Moons,[2] Johannes Reitsma,[2] Maarten van Smeden[2]

[1]Division of Clinical Epidemiology, McGill University Health Centre—Research Institute, Canada
[2]Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Netherlands
[3]Division of Infectious Diseases, McGill University Health Centre, Canada

Correspondence to:
N Dendukuri
nandini.dendukuri@mcgill.ca

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2018;360:j5779
http://dx.doi.org/10.1136/bmj.j5779

Accepted: 20 November 2017

Composite reference standards are used to evaluate the accuracy of a new test in the absence of a perfect reference test. A composite reference standard defines a fixed, transparent rule to classify subjects into disease positive and disease negative groups based on existing imperfect tests. The accuracy of the composite reference standard itself has received limited attention. We show that increasing the number of tests used to define a composite reference standard can worsen its accuracy, leading to underestimation or overestimation of the new test's accuracy. Further, estimates based on composite reference standards vary with disease prevalence, indicating that they may not be comparable across studies. These problems can be attributed to the fact that composite reference standards make a simplistic classification and then ignore the uncertainty in this classification. Latent class models that adjust for the accuracy of the different imperfect tests and the dependence between them should be pursued to make better use of data

SUMMARY POINTS

- Composite reference standards define a fixed, transparent rule to classify subjects into disease positive and disease negative groups based on existing imperfect tests
- They are widely regarded as appropriate for determining sensitivity and specificity of a new test in the absence of a perfect reference test
- Though a composite reference standard is attractive for its simple and transparent construction, it can result in biased estimates as it makes suboptimal use of data
- Bias due to a composite reference standard can worsen as more information is gathered and the new test's accuracy can be overestimated if the errors made by the composite reference standard and the new test are correlated
- Composite reference standards cannot aid standardisation across settings when disease prevalence varies
- Appropriately constructed latent class models should be used to make complete use of the information gathered from multiple imperfect tests

For many diseases, a perfect diagnostic test may not exist or cannot be applied owing to costs or ethical concerns. Researchers evaluating a new test for the disease have no perfect reference against which to compare it. For example, GeneXpert (Xpert) is a nucleic acid amplification test for paediatric pulmonary tuberculosis (TB). Culture is an inadequate reference standard owing to its poor sensitivity,[1] so the accuracy of Xpert is often evaluated by comparing it to a composite reference standard based on multiple imperfect tests.[2 3]

Table 1 shows a composite reference standard defined using data gathered in a South African cohort study of symptomatic children.[6] It classifies a child who is positive on culture, smear microscopy, chest radiography, or the tuberculin skin test as having TB. The apparent advantage of this composite reference standard is that it would identify more TB cases than culture alone. Studies using a standard such as this typically treat it as an error-free reference test to estimate the new test's sensitivity (proportion of all patients with the disease that are correctly detected by the new test) and specificity (proportion of all patients without the disease who are correctly detected by the new test).[3 7] Ensuring the new test is not used to define the composite reference standard is thought to protect against overestimating the new test's accuracy.[8]

Composite reference standards are used for diverse conditions including *Chlamydia trachomatis* infection, apathy, and prostate cancer (see web table 1). Some are required by regulatory authorities for approval of new tests.[9] Despite their widespread use, the number of tests necessary to define an adequate standard is unknown. Composite reference standards based on two or three tests are most common,[7 10] but some are based on upwards of eight or nine tests (web table 1).[2 11]

Consequences of both the composite reference standard and the new test making the same errors are also not well understood. We discuss previously ignored concerns related to composite reference standards using numerical examples and data from a study of childhood TB and draw attention to better methods. We focus on composite reference standards based on the OR rule, which classifies patients with at least one positive test as disease positive and those with all negative tests as disease negative.[3] Other possible composite rules include the AND rule,

Table 1 | A composite reference standard for childhood TB compared with the results of a latent class analysis in a cohort of 749 children

Tests				Composite reference standard*	Xpert	Observed frequency of tests	% children with TB (95% credibility interval) based on latent class analysis	Clinical case definition based on expert consensus[4]*
Culture	Smear	X ray	TST					
–	–	–	–	No TB	+	5	4 (0 to 40)	Unlikely TB
					–	296	2 (0 to 7)	
–	–	–	+	TB	+	7	56 (0 to 100)	Possible TB
					–	149	16 (5 to 33)	
–	–	+	–	TB	+	2	12 (0 to 100)	Possible TB
					–	87	9 (0 to 34)	
–	–	+	+	TB	+	2	88 (50 to 100)	Probable TB
					–	78	52 (26 to 74)	
–	+	–	+	TB	–	1	12 (0 to 100)	Confirmed TB
+	–	–	–	TB	+	1	100 (100 to 100)	Confirmed TB
					–	3	23 (0 to 100)	
+	–	–	+	TB	+	17	100 (100 to 100)	Confirmed TB
					–	8	93 (62 to 100)	
+	–	+	–	TB	+	4	100 (100 to 100)	Confirmed TB
					–	1	54 (0 to 100)	
+	–	+	+	TB	+	27	100 (100 to 100)	Confirmed TB
					–	20	99 (90 to 100)	
+	+	–	–	TB	+	8	100 (100 to 100)	Confirmed TB
+	+	–	+	TB	+	5	100 (100 to 100)	Confirmed TB
+	+	+	–	TB	+	21	100 (100 to 100)	Confirmed TB
+	+	+	+	TB	+	7	100 (100 to 100)	Confirmed TB

TB=pulmonary tuberculosis; TST: tuberculin skin test.
*We defined both the composite reference standard and the clinical case definition using results from culture, smear, chest radiography and TST only. Other composite rules for defining TB have been published.[2 5] A more recent case definition includes Xpert and may give a slightly different classification.[1]

which classifies a patient as disease positive only if all tests are positive, or K positive rules, which classify a patient as disease positive only if at least K tests are positive.[12]

Numerical examples

In these examples, the sensitivities of the component tests used to define the composite reference standard are moderate to high and their specificities are near perfect. An OR rule composite reference standard is, therefore, anticipated to have higher sensitivity than a single imperfect reference test.

We generated data for a sample of 1000 people assuming the composite reference standard was made up of component tests with sensitivity of 70% and specificities of 98-100% (detailed explanation in the web appendix). Disease prevalence was assumed to be 10%—that is, 100 patients were disease positive and 900 were disease negative. The new test under evaluation was set to have a sensitivity of 90% and a specificity of 98%.

Increasing the number of component tests

Misclassification of disease status

Depending on their accuracy, adding more component tests to the composite reference standard might cause more misclassification rather than less. We start with the ideal situation where each component test has perfect specificity of 100% and where the different component tests are conditionally independent (meaning that they are not prone to making the same false positive or false negative errors). Table 2 lists the frequency of results on the component tests and the composite reference standards based on them. As we move from a single reference test to a composite reference standard based on two or three component tests, the number of patients correctly classified as having the disease increases from 70 (34+15+15+6) to 91 (34+15+15+6+15+6) to 97 (34+15+15+6+15+6+6). The gain at the second step is less than at the first. After about five component tests (data not shown), additional tests increase costs but don't result in any gain, as all 1000 patients are correctly identified.

Table 2 shows how misclassification changes when the specificity of the component tests decreases to 98%. The classification of disease positive patients remains the same, but the number of misclassified disease negative patients increases from 17 to 51 as we move from a single reference test to a composite reference standard with three component tests. For this example, the composite reference standard with three component tests resulted in more misclassified patients overall (three false positives and 51 false negatives) than the single reference test (30 false positives and 17 false negatives). The overall number of misclassified patients will continue to rise as more component tests are added to the composite reference standard.[12]

Sensitivity and specificity of new tests

Using the data in table 2 we can show that when all component tests have perfect specificity and are conditionally independent, the sensitivity of the new test is estimated accurately at its value of 90% irrespective of the number of tests in the composite reference standard (fig 1a). The estimate of the specificity of the new test steadily improves with every added test until it reaches the true value (fig 1b).

doi: 10.1136/bmj.j5779 | *BMJ* 2018;360:j5779 | the bmj

Table 2 | Expected frequency of results on individual tests and OR rule based composite reference standards in 1000 subjects*

True disease status	Results of component tests in composite reference standard (T1, T2, T3)	Single reference test (T1)	Composite reference based on T1, T2	Composite reference based on T1, T2, T3	Frequency of results under perfect specificity	Frequency under specificity of 98%
+	+ + +	+	+	+	34	34
+	+ + −	+	+	+	15	15
+	+ − +	+	+	+	15	15
+	+ − −	+	+	+	6	6
+	− + +	[−]	+	+	15	15
+	− + −	[−]	+	+	6	6
+	− − +	[−]	[−]	+	6	6
+	− − −	[−]	[−]	[−]	3	3
−	+ + +	[+]	[+]	[+]	0	0
−	+ + −	[+]	[+]	[+]	0	0
−	+ − +	[+]	[+]	[+]	0	0
−	+ − −	[+]	[+]	[+]	0	17
−	− + +	−	[+]	[+]	0	0
−	− + −	−	[+]	[+]	0	17
−	− − +	−	−	[+]	0	17
−	− − −	−	−	−	900	849

*Tests are conditionally independent, each having a sensitivity of 70%. True disease prevalence is 10%. Square brackets indicate misclassification of the true disease status by the single reference test or a composite reference standard. The single reference test (T1) picks up 34+15+15+6=70 true disease cases accurately. The composite reference standard based on T1 and T2 adds 15+6=21 to this total. The composite reference standard based on T1, T2 and T3 picks up a further 6 cases. When the specificity of the component tests is 100%, no disease negative cases are misclassified. When the specificity drops to 98%, the number of disease negative cases misclassified by the reference increases from 17 to 34 to 51 with each added test.

When the specificity of the component tests falls to 98%, the specificity estimates of the new test are almost identical to those obtained previously, but the sensitivity of the new test is now underestimated (fig 1). When the composite reference standard was composed of three tests, for example, the estimated sensitivity was 59%, much lower than the true value of 90%. This underestimation worsens with every test added to the composite reference standard.

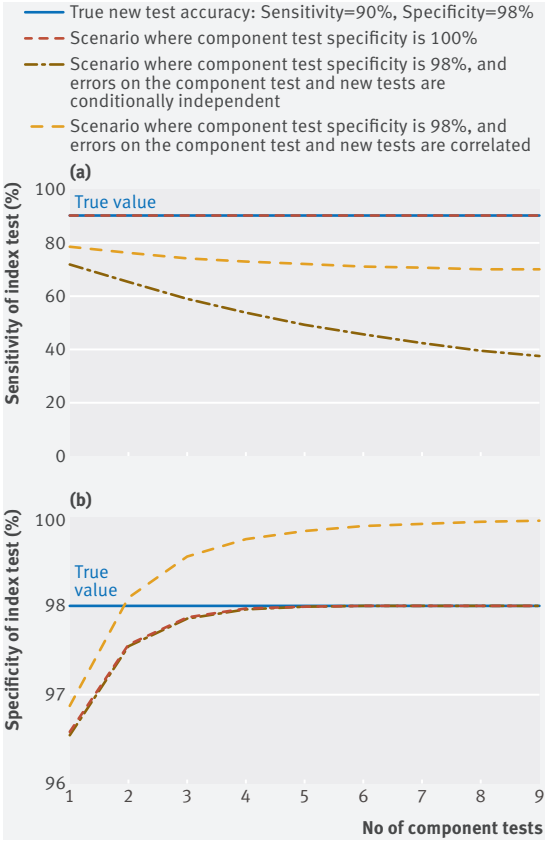

Fig 1 | Change in estimates of sensitivity and specificity of the new test with increasing number of component tests in the composite reference standard.

Overestimating sensitivity or specificity of the new test

So far, we have assumed that all tests—component tests in the composite reference standard and the new test—are conditionally independent. In practice, however, errors made by multiple tests might be correlated.[13] In studies evaluating new tests for *Chlamydia trachomatis*, for example, the component tests and the new test are typically nucleic acid amplification tests, which risk making the same false positive error of detecting a non-viable organism.[14] In these situations, the composite reference standard can overestimate the accuracy of the new test because it does not adjust for the presence of conditional dependence—that is, the errors of the new test remain undetected.

To study the effects of conditional dependence, we generated data from a setting where both tests are likely to make the same false positive errors even though their specificity remains high at 98% (see web table 2 for details).[6] As in our previous example, the sensitivity of the new test is underestimated, and this worsens with each component test added to the composite reference standard (fig 1). The specificity of the new test is underestimated when compared to a single imperfect test but becomes overestimated as component tests are added to the composite reference standard (fig 1). As the number of component tests increases, the estimated specificity of the new test converges to a value higher than the true value.[12] In our example, the new test's specificity will converge at 99.94%, compared with its true specificity of 98%. This may not seem like a large magnitude of bias but could lead to an important underestimate of the number of false positives the new test will produce in a low prevalence population.

Comparability across studies

When a new test is compared to the same composite reference standard in two different studies, the reported value of the new test's accuracy will depend on the disease prevalence in each study. Using a standardised composite reference standard therefore does not ensure comparability across studies. We considered a composite reference standard based on three conditionally independent component tests, each with sensitivity 70% and specificity 98%. Disease prevalence ranged from low (5%) to high (30%), as might be expected across geographic regions or healthcare settings. The new test's sensitivity was assumed to be 90% and its specificity 98%, as before. We found the estimated sensitivity of the new test ranged from 43% when the prevalence was 5% to 79% when the prevalence is 30% (web figure 1). Estimated specificity did not vary greatly with prevalence for the settings we used.

Latent class models can make better use of data

The drawbacks of the composite reference standard can be overcome using a statistical modelling approach called latent class analysis.[3] Instead of classifying subjects into fixed disease categories, latent class analysis estimates the probability that each patient has the disease using all observed tests, including the test under evaluation (web figure 2). It adjusts for the sensitivity and specificity of each test as well as the possibility of conditional dependence between them. Simply put, latent class analysis considers how certain we are about classifying patients into diseased or non-diseased groups rather than making a black and white decision. Column 8 of table 1 shows the estimated risk of TB for each observed combination of tests based on a recent latent class analysis for the childhood TB data.[6]

Notably, the estimated risk of TB from this latent class analysis follows the gradation proposed by an expert group's clinical case definition (column 9 of table 1),[4] which classifies subjects into confirmed TB, probable TB, possible TB, and unlikely TB groups, using the same four tests as the composite reference standard. Our data show that composite reference standard would classify confirmed, probable and possible TB cases all as having TB, resulting in an estimated prevalence of 94%. Using culture as a reference would only consider the confirmed TB cases, resulting in an underestimate of the prevalence (16.4%). Latent class analysis estimates a 100% risk of TB for most cases of confirmed TB, though the risk is lower for unusual patterns. The risk of TB among the probable and possible TB cases ranges from 9% to 52% among patients with a negative Xpert test but increases when Xpert is positive. The resulting prevalence estimate based on latent class analysis is 26.7%. Because the latent class analysis adjusts for conditional dependence between culture and Xpert, it also provides a more realistic estimate of Xpert sensitivity (49.4%) than would be obtained with culture (74.4%) or the composite reference standard (22.5%) (see web material for how the latent class analysis estimates were calculated).

The advantages of latent class analysis are accompanied by the challenges of using a more sophisticated analytical technique. Construction of these models requires interdisciplinary expertise of both methodologists and clinicians[6] to determine the particular tests, covariates, conditional dependence structure, and previous knowledge to be considered. Validation of these models against a perfect reference may not always be possible. Sometimes competing models cannot be distinguished using standard statistical methods.[15] This is not a drawback of latent class analysis, but a reflection of the uncertainty in our knowledge due to the lack of a perfect reference test. Comparison with external information, such as the experts' clinical case definition, can aid in assessing whether the model provides sensible results. This step is important because, as with all statistical models, incorrect model specification can lead to biased results.[16]

Discussion

Composite reference standards are considered valid for estimating diagnostic accuracy when no perfect reference standard exists.[3 7] But we have shown that the OR rule based composite reference standard leads to biased estimates of the accuracy of a new test unless stringent conditions hold. The additional information gathered from each component test results in worsening bias. Our previous work has shown these observations also apply to composite reference standards based on the AND rule and or the K positive rule.[12]

Composite reference standards may be considered clinically meaningful[17] as they resemble clinical decision rules, which classify patients into mutually exclusive categories to support decision making—for example, rules identifying whether a subject is a candidate for TB treatment. Such decision rules are not necessary in research settings as no black or white decision needs to be made. Clinical decision rules might indicate the best possible management strategy, but are recognised by clinicians as imperfect.[16] Yet similar rules are used to define composite reference standards for a diagnostic accuracy studies with no such recognition.

In the absence of a perfect reference test, a new test could be evaluated in terms of outcomes such as diagnostic yield or effect on patient management instead of accuracy.[18] Latent class analysis would also be relevant in such analyses to estimate percentage of overdiagnosis or overtreatment,[6 19] eventually supporting the development of optimal clinical decision rules.

As our ability to measure results on multiple tests/biomarkers increases, development of appropriate latent class models should be pursued in the absence of a perfect reference test to make optimal use of the data gathered.

Linked information

- More detail on the problems with composite reference standards https://www.ncbi.nlm.nih.gov/pubmed/26555849

doi: 10.1136/bmj.j5779 | *BMJ* 2018;360:j5779 | the bmj

- A review paper on use of latent class models for diagnostic research https://www.ncbi.nlm.nih.gov/pubmed/24272278
- Applications of latent class models with free accompanying software https://www.ncbi.nlm.nih.gov/pubmed/27737841, https://www.ncbi.nlm.nih.gov/pubmed/7840100
- Guidelines for reporting latent class models https://www.equator-network.org/reporting-guidelines/stard-blcm/

Contributors and sources: The authors include biostatisticians, epidemiologists, and clinicians with expertise in diagnostic research and a particular interest in methods for evaluating diagnostic accuracy in the absence of a perfect reference test. All authors participated in planning and writing the paper. IS generated the numerical examples. ND is the guarantor.

Funding: This work was supported by funding from the Canadian Institutes of Health Research (Grant number 89857).

Competing interests: All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/coi_disclosure.pdf and declare support from the Canadian Institutes of Health Research for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

1 Graham SM, Cuevas LE, Jean-Philippe P, et al. Clinical case definitions for classification of intrathoracic tuberculosis in children: an update. *Clin Infect Dis* 2015;61(Suppl 3):S179-87. doi:10.1093/cid/civ581
2 Yin Q-Q, Jiao W-W, Han R, et al. Rapid diagnosis of childhood pulmonary tuberculosis by Xpert MTB/RIF assay using bronchoalveolar lavage fluid. *Biomed Res Int* 2014;2014:310194. doi:10.1155/2014/310194
3 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. *J Clin Epidemiol* 2009;62:797-806. doi:10.1016/j.jclinepi.2009.02.005
4 Graham SM, Ahmed T, Amanullah F, et al. Evaluation of tuberculosis diagnostics in children: 1. Proposed clinical case definitions for classification of intrathoracic tuberculosis disease. Consensus from an expert panel. *J Infect Dis* 2012;205(Suppl 2):S199-208. doi:10.1093/infdis/jis008
5 Hatherill M, Hanslo M, Hawkridge T, et al. Structured approaches for the screening and diagnosis of childhood tuberculosis in a high prevalence region of South Africa. *Bull World Health Organ* 2010;88:312-20. doi:10.2471/BLT.09.062893
6 Schumacher SG, van Smeden M, Dendukuri N, et al. Diagnostic test accuracy in childhood pulmonary tuberculosis: a bayesian latent class analysis. *Am J Epidemiol* 2016;184:690-700. doi:10.1093/aje/kww094
7 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. doi:10.1002/(SICI)1097-0258(19991130)18:22<2987::AID-SIM205>3.0.CO;2-B
8 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;15:106-14. doi:10.1128/CVI.00223-07
9 US Food and Drug Administration Draft guidance for infectious disease next generation sequencing based diagnostic devices: microbial identification and detection of antimicrobial resistance and virulence markers. 2016. https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM500441.pdf.
10 Black CM, Marrazzo J, Johnson RE, et al. 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. *J Clin Microbiol* 2002;40:3757-63. doi:10.1128/JCM.40.10.3757-3763.2002
11 Shrier LA, Dean D, Klein E, Harter K, Rice PA. Limitations of screening tests for the detection of Chlamydia trachomatis in asymptomatic adolescent and young adult women. *Am J Obstet Gynecol* 2004;190:654-62. doi:10.1016/j.ajog.2003.09.063
12 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. doi:10.1002/sim.6803
13 Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance[pii]. *Stat Med* 1997;16:2157-75. doi:10.1002/(SICI)1097-0258(19971015)16:19<2157::AID-SIM653>3.0.CO;2-X
14 Hadgu A, Dendukuri N, Hilden J. Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. *Epidemiology* 2005;16:604-12. doi:10.1097/01.ede.0000173042.07579.17
15 Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 2004;60:427-35. doi:10.1111/j.0006-341X.2004.00187.x
16 Green SM, Schriger DL, Yealy DM. Methodologic standards for interpreting clinical decision rules in emergency medicine: 2014 update. *Ann Emerg Med* 2014;64:286-91. doi:10.1016/j.annemergmed.2014.01.016
17 Pepe MS. *The statistical evaluation of medical tests for classification and prediction* Oxford Statistical Science Series, 2006.
18 Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ* 2012;344:e686. doi:10.1136/bmj.e686
19 Xie X, Sinclair A, Dendukuri N. Evaluating the accuracy and economic value of a new test in the absence of a perfect reference test. *Res Synth Methods* 2017;8:321-32. doi:10.1002/jrsm.1243

Web appendix: Calculations, web tables and figures