BMJ

RESEARCH METHODS & REPORTING

Value of composite reference standards in diagnostic research

Combining several tests is a common way to improve the final classification of disease status in diagnostic accuracy studies but is often used ambiguously. This article gives advice on proper use and reporting of composite reference standards

Christiana A Naaktgeboren *PhD fellow*, Loes C M Bertens *PhD fellow*, Maarten van Smeden *PhD fellow*, Joris A H de Groot *assistant professor*, Karel G M Moons *professor*, Johannes B Reitsma *associate professor*

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, Netherlands

A common challenge in diagnostic studies is to obtain a correct final diagnosis in all participants. Ideally, a single error-free reference test, known as a gold standard, is used to determine the final diagnosis[1] and estimate the accuracy of the test or diagnostic model under evaluation. If the reference standard does not perfectly correspond to true target disease status, estimates of the accuracy of the test or model under study (index test), such as sensitivity, specificity, predictive values, or area under the curve, can be biased.[2] This is known as imperfect reference standard bias. One method to reduce this bias is to use a fixed rule to combine results of several imperfect tests into a composite reference standard.[3] When the combination of several component tests provides a better perspective on disease than any of the individual tests alone, accuracy estimates of the test under evaluation (the index test) will be less biased than if only one imperfect test is used as the reference standard. Comparing the index test against each component test separately and then averaging the accuracy estimates is not recommended; it is better to insightfully combine component tests together into a composite reference standard.

The hallmark of composite reference standards is that each combination of test results leads to a particular final diagnosis; in its simplest form, disease present or absent. For example, in a study on the accuracy of a rapid antigen test for detecting trichomoniasis, researchers decided against using the traditional gold standard of culture because it probably misses some cases.[4] As they believed that microscopy picks up additional true cases, they instead considered patients as diseased if either microscopy or culture results were abnormal. Table 1⇓ gives further examples.

Although the choice of component tests and the rules used to combine them affects the estimates of accuracy of the test under study,[7] little guidance exists on how to develop and define a composite reference standard. Additionally, there is a lack of consensus in the way the term composite reference standard is used and reporting of results is generally poor. To address these problems, we provide an explanation of the methods for composite reference standards and make recommendations for development and reporting.

What is a composite reference standard?

A composite reference standard is a fixed rule used to make a final diagnosis based on the results of two or more tests, referred to as component tests. For each possible pattern of component test results (test profiles), a decision is made about whether it reflects presence or absence of the target disease.

Composite reference standards are appealing because of their similarity to clinical practice; they strongly resemble diagnostic rules that exist for several conditions, such as rheumatic fever and depression. Their main advantage is reproducibility of results, which is made possible by the transparency and consistency in the way that the final diagnosis is reached across participants. However, they also have disadvantages, the most glaring being the subjectivity introduced in the development of the rule.

The term "composite reference standard" is often loosely used as a catch-all term to describe any situation in which multiple reference tests are used to evaluate the accuracy of the index test. It is sometimes mistakenly used to describe differential verification, when different reference standards are used for different groups of participants (table 2⇓).[8 9] It has also been used to describe discrepant analysis, a method in which the reference standard is re-run or re-evaluated, or a different reference standard is used, when the first one does not agree with the index test.[13] Both these approaches can lead to seriously biased estimates of accuracy and should be avoided whenever possible.

In the example in table 2⇓ of a study on deep venous thrombosis differential verification was mislabelled as a composite reference standard. The reference standard for participants with a negative

Correspondence to: C A Naaktgeboren c.naaktgeboren@umcutrecht.nl

index test result was clinical follow-up while those with a positive result received the preferred reference standard, computed tomography.[11] If minor thromboembolisms that would have been picked up by computed tomography were missed during follow-up, the number of false negatives will be underestimated and the number of true negatives overestimated, thus biasing the accuracy estimates. Ethical or practical difficulties sometimes make it impossible to implement the same reference standard in all participants, but it is important that the term differential verification is used to describe such situations.

Table 2⇓ also gives an example of discrepant analysis from an imaging study for coronary artery stenosis in which the reference standard results were re-evaluated when they did not agree with the index test results.[12] Such re-evaluation can only lead to increased agreement between index test and the reference standard, which in turn can only lead to overestimates of accuracy. Although discrepant analysis his highly discouraged, situations in which the reference standard is repeated or a different reference standard is applied in those patients where the index test and first reference standard disagree, should be termed discrepant analysis.

To avoid confusion we recommend using the term composite reference standard exclusively for situations in which, by design, all patients are intended to receive the same component tests and these component tests are interpreted and combined in a fixed way for all patients.

Developing a composite reference standard

As the choice of component tests and the rule for combining them strongly influences the accuracy of composite reference standards,[14] careful attention is required when developing the decision rule. Ideally, the combination of test results and the corresponding final diagnosis should be specified before the study to prevent data driven decisions. However, if there is uncertainty about the best composite reference standard, a sensitivity analysis could be planned to see how sensitive the results are to the particular choice of tests or combination rule. It is also important that the composite reference standard is clinically relevant. In other words, it should detect cases that will benefit from clinical intervention rather than simply the presence of disease.[15] For clinical situations when the true disease status cannot be defined the composite reference standard should reflect the provisional working definition. Keeping diagnostic guidelines in mind and seeking advice from experts in the field will help ensure that the chosen standard is clinically relevant and interpretable.

Defining rules to combine component tests

Two rules exist for combining component tests into a composite reference standard. In the simplest scenario of two dichotomous component tests, participants could be considered to have the disease if either test is indicative of disease (any positive rule, also known as the “or” rule). The alternative is that participants are considered to have the disease only if both tests detect disease (all positive or “and” rule). If there are more than two component tests a combination of these two rules can be used.

Increasing the number of component tests will increase the number of participants categorised as diseased. If the any positive rule is used, this will increase the sensitivity of the composite reference standard (more diseased subjects will be classified as diseased) but decrease its specificity (more non-diseased subjects will be classified as having the disease). The reverse is true for the “all positive” rule; sensitivity of the composite reference standard decreases while specificity increases. Table 3⇓ gives an example of how the choice of combination rule affects the accuracy of the composite reference standard, which in turn affects the accuracy estimates of the test under study.[2]

There is almost always a trade-off between sensitivity and specificity when considering alternative ways to combine component tests.[14] The exception is when a component test in an “any positive” rule has perfect sensitivity, which makes a composite reference standard with perfect sensitivity, or when a component test in an “all positive” rule has perfect specificity, which makes a composite standard with perfect specificity.[3] Near perfect sensitivity or specificity of a component test is often the reasoning provided for the rule chosen.

Selection of component tests

Although it may be tempting to include numerous component tests, the gain in sensitivity or specificity of the resulting composite reference standard decreases (and the clinical interpretability may diminish) as more tests are added. This is because additional tests may fail to provide new information. In the trichomoniasis example, if another test such as polymerase chain reaction amplification is added, new true cases may be detected.[4] However, if yet another test is added, fewer additional true cases will be detected because fewer remained undetected. Eventually, all true cases are detected and additional tests will only result in false positive results, thus decreasing the specificity of the composite reference standard.

Multiple tests will be useful only if the component tests catch each other’s mistakes. For example, in a group of patients who truly have trichomoniasis, if microscopy identifies disease in the same participants as culture does, microscopy does not add any information and therefore the sensitivity of the composite reference standard will not be higher than that of culture alone.[2] When component tests make the same classifications in truly diseased or non-diseased patients more or less often than is expected by chance alone, this is referred to as conditional dependence.

In some cases, conditional dependence can be avoided or reduced by choosing component tests that look at different biological aspects of the disease.[16] To avoid causing the tests to make the same mistakes, you should consider blinding the observer of each component test to the results of the other component tests if knowledge of these other test results can influence interpretation.

Extensions to the basic composite reference standard

The basic composite reference standard categorises patients simply as diseased or non-diseased. However, multiple disease categories can also be defined, such as subtypes, stages, or degree of certainty of disease. An example is a study on tuberculosis in which people were categorised into one of four levels of disease certainty (table 4⇓).[17]

The basic composite reference standard gives equal weight to all tests, but in clinical practice tests carry different weights. The relative importance of the component tests can be incorporated by assigning weights. For example, in the assessment of adherence to isoniazid treatment for latent tuberculosis in table 1⇓, the most reliable test was given twice the weight of the other tests.[6]

Missing values on component tests

As with all diagnostic accuracy studies, results may be biased when not all participants receive the intended reference standard.[8] Careful attention needs to be paid to missing values in component tests. For example, if the “any positive” rule is used and the result of component test 1 is positive, we can conclude that a patient is diseased without knowing the result of component test 2. For efficiency, researchers might consider skipping the second test in participants whose first test result is positive.[3 18] However, if component test 1 is negative, component test 2 becomes necessary for determining the diagnosis.

When a result is missing from a component test that must be present under the combination rules, the composite reference standard is also missing. This may affect the accuracy estimates of the index test and mathematical methods should be used to tentatively correct for this bias.[19]

Reporting guidelines

Complete and accurate reporting of the reference standard procedure is critical to allow readers to judge the potential risk of bias in accuracy estimates. This is especially important for systematic reviews of diagnostic tests. The validity of comparing accuracy estimates between studies and pooling of estimates across studies is challenged when studies use different reference standards or when reference standards are poorly defined or reported.[20 21] We therefore recommend that in addition to using current reporting guidelines,[22] authors of diagnostic accuracy studies should include the following details about studies with composite reference standards:

- The rationale behind the selection of component tests and the combination rule
- The corresponding final diagnosis for each combination of test results
- Whether component test results were missing and and whether this resulted in a missing composite reference standard
- The number of participants with each combination of test results. For continuous tests, this information should at least be provided for the optimal or most common cut-off point.

Table 5⇓ gives a template for reporting. The availability of all of the above information will allow studies using composite reference standards to be compared with those using only one of the component tests as the reference standard.

Conclusions and recommendations

Combining multiple tests to define a target disease status rather than using a single imperfect test is a transparent and reproducible method for dealing with the common problem of imperfect reference standard bias. Although composite reference standards may reduce the amount of such bias, they cannot completely eliminate it because it is unlikely that a combination of imperfect tests will produce a composite standard with perfect sensitivity and specificity.

Other methods for dealing with bias resulting from imperfect reference standards are panel diagnosis and latent class analysis.[1 3] In panel diagnosis, multiple experts review relevant patient characteristics, test results, and sometimes follow-up information before coming to a consensus about the final diagnosis in each patient. Latent class analysis estimates accuracy by assuming that true disease status is unobservable and relating the results of multiple tests to it in a statistical model.[3 23] The choice of method to deal with imperfect reference standard bias will probably depend on the type, number, and accuracy of the pieces of diagnostic information available in a particular study. Results from all three methods could be presented to strengthen their face validity. Researchers who use a composite reference standard can improve the transparency and reproducibility of their results by following our recommendations on reporting.

Contributors: All authors participated in the conception and design of the article, worked on the drafting of the article and revising it critically for important intellectual content, and have approved the final version to be published. CAN had the idea for the article, performed the literature search, and wrote the article. JBR is the guarantor.

Competing interests: All authors read and understood the BMJ policy on declaration of interests and declare financial support from Netherlands Organization for Scientific Research (project 918.10.615).

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007;11:iii, ix-51.
2 Walter SD, Macaskill P, Lord SJ, Irwig L. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Stat Med* 2012;31:1129-38.
3 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.
4 Hegazy MM, El-Tantawy NL, Soliman MM, El-Sadeek ES, El-Nagar HS. Performance of rapid immunochromatographic assay in the diagnosis of Trichomoniasis vaginalis. *Diagn Microbiol Infect Dis* 2012;74:49-53.
5 Siba V, Horwood PF, Vanuga K, Wapling J, Sehuko R, Siba PM, et al. Evaluation of serological diagnostic tests for typhoid fever in Papua New Guinea using a composite reference standard. *Clin Vaccine Immunol* 2012;19:1833-7.
6 Nicolau I, Tian L, Menzies D, Ostiguy G, Pai M. Point-of-care urine tests for smoking status and isoniazid treatment monitoring in adult patients. *PLoS One* 2012;7:e45913.
7 Hadgu A, Dendukuri N, Wang L. Evaluation of screening tests for detecting Chlamydia trachomatis: bias associated with the patient-infected-status algorithm. *Epidemiology* 2012;23:72-82.
8 De Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. *BMJ* 2011;343:d4770.
9 Naaktgeboren CA, de Groot JAH, van Smeeden M, Moons KGM, Reitsma JB. Evaluating diagnostic accuracy in the face of multiple reference standards. *Ann Intern Med* 2013;159:195-202.
10 Ewer AK, Furmston AT, Middleton LJ, Deeks JJ, Daniels JP, Pattison HM, et al. Pulse oximetry as a screening test for congenital heart defects in newborn infants: a test accuracy study with evaluation of acceptability and cost-effectiveness. *Health Technol Assess* 2012;16:v-184.
11 Geersing GJ, Erkens PM, Lucassen WA, Buller HR, Cate HT, Hoes AW, et al. Safe exclusion of pulmonary embolism using the Wells rule and qualitative D-dimer testing in primary care: prospective cohort study. *BMJ* 2012;345:e6564.
12 Kerl JM, Schoepf UJ, Zwerner PL, Bauer RW, Abro JA, Thilo C, et al. Accuracy of coronary artery stenosis detection with CT versus conventional coronary angiography compared with composite findings from both tests as an enhanced reference standard. *Eur Radiol* 2011;21:1895-903.
13 Hadgu A. The discrepancy in discrepant analysis. *Lancet* 1996;348:592-3.
14 Macaskill P, Walter SD, Irwig L, Franco EL. Assessing the gain in diagnostic performance when combining two diagnostic tests. *Stat Med* 2002;21:2527-46.
15 Lord SJ, Staub LP, Bossuyt PM, Irwig LM. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. *BMJ* 2011;343:d4684.
16 Gardner IA, Stryhn H, Lind P, Collins MT. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Prev Vet Med* 2000;45:107-22.
17 Vadwai V, Boehme C, Nabeta P, Shetty A, Alland D, Rodrigues C. Xpert MTB/RIF: a new pillar in diagnosis of extrapulmonary tuberculosis? *J Clin Microbiol* 2011;49:2540-5.
18 Hilden J. Boolean algebra, Boolean nodes. In: Kattan M, Cowen ME, eds. Encyclopedia of medical decision making. 1st ed. Sage, 2009:94-8.
19 De Groot JA, Janssen KJ, Zwinderman AH, Bossuyt PM, Reitsma JB, Moons KG. Correcting for partial verification bias: a comparison of methods. *Ann Epidemiol* 2011;21:139-48.
20 Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529-36.
21 Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21:1525-37.
22 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. *Fam Pract* 2004;21:4-10.
23 Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2007;8:474-84.

Accepted: 11 September 2013

Cite this as: *BMJ* 2013;347:f5605

Summary points

- A composite reference standard is a predefined rule that combines the results of multiple imperfect (component) tests in order to improve the classification of disease status in a diagnostic study
- The term is often misused to describe differential verification, a situation in which different reference standards are used for different groups of participants
- Different sets of component tests or different rules to combine the same component tests will lead to different estimates of accuracy for the test(s) under study
- When using composite reference standards, it is important to prespecify and explain the rationale for the rule, report index test results for each combination of component tests, and explain how missing component test results are dealt with

Tables

Table 1| **Examples of composite reference standards**

Condition	Example	Rule for combination
Trichomoniasis[4]	"Samples were labeled as positive if the results of either mount microscopy or culture were positive… samples were labeled negative if both mount preparations and culture were negative"	Any positive rule
Typhoid fever[5]	"A composite reference standard of blood culture and polymerase chain reaction was used"	Any positive rule
Adherence to isoniazid preventive therapy for latent tuberculosis[6]	Adherence defined as ≥3 points when tests receive the following weights:	Heavier weights given to more accurate tests
	2 points for a positive urine isoniazid test result	
	1 point for patient observed taking tablets	
	1 for hospital records	
	1 point for patient self reporting	

Table 2| **Examples of misuse of the term composite reference standard**

Disease	Example	Explanation of misuse
Congenital heart defect[10]	"Pulse oximetry was performed prior to discharge and the results of this index test were compared with a composite reference standard (echocardiography, clinical follow-up and follow-up through interrogation of clinical databases)."	This is differential verification because some patients received an intensive clinical work-up while others were followed-up in clinical databases
Deep venous thrombosis[11]	"All patients were… diagnosed according to local protocols. Pulmonary embolism was confirmed or refuted on the basis of a composite reference standard, including spiral computed tomography and three months' follow-up."	This is differential verification because high risk patients had computed tomography whereas other patients were followed- up
Coronary artery stenosis[12]	"Diagnosis stenosis using composite findings from both [the index and the reference] tests as an enhanced reference standard . . . If a stenosis ≥50% had been seen on one [imaging test] but not on the other test, the observers closely re-evaluated the respective coronary artery segment showing discordant findings in order to confirm or revise their initial interpretation."	This is an example of discrepant analysis[13] in which the index test influences the reference standard result

Table 3| **Effect of using different rules to produce composite reference standard on estimates of accuracy using example inspired by a study on the accuracy of rapid antigen detection test for trichomoniasis**[4]

Result of component reference tests		Diagnosis with composite reference standard		Index test (rapid antigen detection test, n=100)	
Culture	Microscopy	Any positive rule*	All positive rule†	No with positive result	No with negative result
+	+	+	+	25	1
+	–	+	–	10	3
–	+	+	–	4	1
–	–	–	–	1	55

*Accuracy estimate using the any positive rule: sensitivity=(25+10+4)/((25+10+4)+(1+3+1))=0.89; specificity=55/(55+1)=0.98.

†Accuracy estimate using the all positive rule: sensitivity=25/(25+1)=0.96; specificity=(3+1+55)/((3+1+55)+(10+4+1))=0.8.

Table 4| Use of a composite reference standard to determine different categories of diagnosis for turberculosis[17]

	Individual tests				
Final diagnosis	Acid fast bacilli smear	Culture	Radiology	Histology	Follow-up
Confirmed	+/–	+	+/–	+/–	+
Probable	+/–	–	+	+	+
	+/–	–	+	–	+
	+/–	–	–	+	+
Possible	+/–	–	–	–	+
Not tuberculosis	–	–	–	–	–

Table 5| **Template for reporting results when using a composite reference standard**

Composite reference standard				Index test*	
Test 1	Test 2	Test 3	Final diagnosis	No with positive result	No with negative result
+	+	+	+	p_1	n_1
+	+	–	+ or –	p_2	n_2
+	–	+	+ or –	p_3	n_3
+	–	–	+ or –	p_4	n_4
–	+	+	+ or –	p_5	n_5
–	+	–	+ or –	p_6	n_6
–	–	+	+ or –	p_7	n_7
–	–	–	–	p_8	n_8