Chapter 9: Options for Summarizing Medical Test Performance in the Absence of a "Gold Standard"

Thomas A. Trikalinos, MD[1,2] and Cynthia M. Balion, PhD[3,4]

[1]Center for Evidence-based Medicine, and Department of Health Service, Policy and Practice, Brown University, Providence, RI, USA; [2]Tufts Evidence-based Practice Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA; [3]Department of Pathology and Molecular Medicine, Hamilton General Hospital, Hamilton, ON, Canada; [4]McMaster Evidence-based Practice Center, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada.

The classical paradigm for evaluating test performance compares the results of an index test with a reference test. When the reference test does not mirror the "truth" adequately well (e.g. is an "imperfect" reference standard), the typical ("naïve") estimates of sensitivity and specificity are biased. One has at least four options when performing a systematic review of test performance when the reference standard is "imperfect": (a) to forgo the classical paradigm and assess the index test's ability to predict patient relevant outcomes instead of test accuracy (i.e., treat the index test as a predictive instrument); (b) to assess whether the results of the two tests (index and reference) agree or disagree (i.e., treat them as two alternative measurement methods); (c) to calculate "naïve" estimates of the index test's sensitivity and specificity from each study included in the review and discuss in which direction they are biased; (d) mathematically adjust the "naïve" estimates of sensitivity and specificity of the index test to account for the imperfect reference standard. We discuss these options and illustrate some of them through examples.

KEY WORDS: medical test; diagnostic test; alloy standard.
J Gen Intern Med 27(Suppl 1):S67–75
DOI: 10.1007/s11606-012-2031-7

INTRODUCTION

In the classical paradigm for evaluating the "accuracy" or performance of a medical test (index test), the results of the test are compared with the "true" status of every tested individual or every tested specimen. Sometimes, this "true" status is directly observable (e.g., for tests predicting short-term mortality after a procedure). However, in many cases the "true" status of the tested subject is judged based upon another test as a reference method. Problems can arise when the reference test does not mirror the "truth" adequately well: one will be measuring the performance of the index test against a faulty standard, and is bound to err. In fact, the worse the deviation of the reference test from the unobserved "truth", the poorer the estimate of the index test's performance will be. This is otherwise known as "reference standard bias".[1–4]

In this paper we discuss how researchers engaged in the Effective Healthcare Program of the United States Agency for Healthcare Research and Quality (AHRQ) think about synthesizing data on the performance of medical tests when the reference standard is "imperfect". Because this challenge is a general one and not specific to AHRQ's program, we anticipate that the current paper is of interest to the wider group of those who perform or use systematic reviews of medical tests. Of the many challenges that pertain to issues with reference standards, we will discuss only one, namely, the case of a reference standard test that itself misclassifies the test subjects at a rate we are not willing to ignore ("imperfect reference standard"). We will not discuss verification bias, where the use of the reference standard is guided by the results of the index test and is not universal.

IMPERFECT REFERENCE STANDARDS

What is Meant by "Imperfect Reference Standard" and Why is it Important for Meta-Analysis and Synthesis in General?

Perhaps the simplest case of test performance evaluation includes an "index test" and a reference test ("reference standard") whose results are dichotomous in nature (or are made dichotomous). Both tests are used to inform on the presence or absence of the condition of interest, or predict the occurrence of a future event. For the vast majority of medical tests, both the results of the index test and the reference test can be different than the true status of the condition of interest. Figure 1 shows the correspondence between the true 2 × 2 table probabilities (proportions) and the eight strata defined by the combinations of index and reference test results and the presence or absence of the condition of interest. These 8 probabilities (α_1, β_1, γ_1, δ_1, α_2, β_2, γ_2 and δ_2) are not known, and have to be estimated from the data (from studies of diagnostic or prognostic

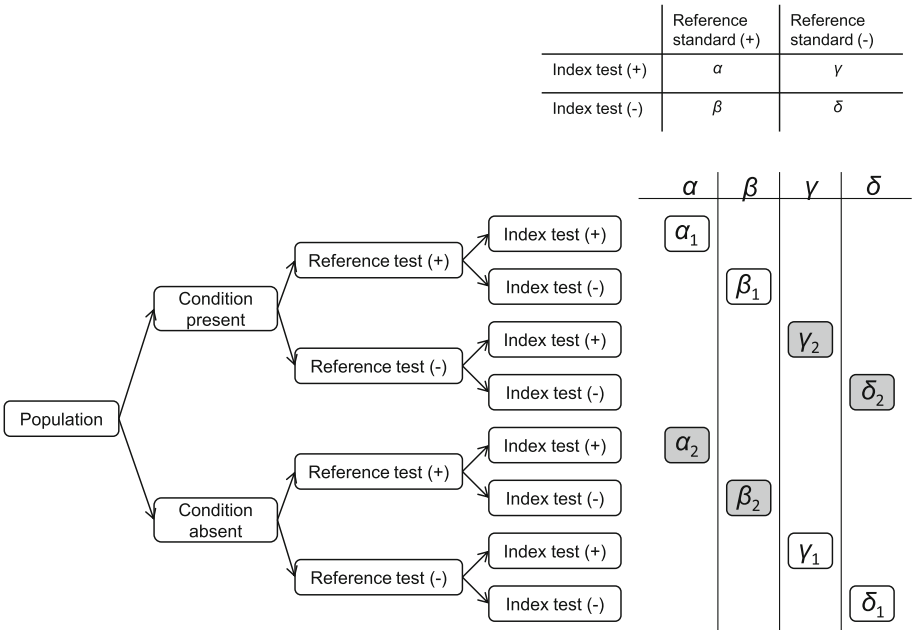

Figure 1. Correspondence of test results and true proportions in the 2 × 2 table. The cells in 2 × 2 the table, α, β, γ, δ are the true population proportions corresponding to combinations of test results. The diagram depicts how these proportions break down according to the (unknown) true status of the condition of interest. For example, the proportion when both the index test and the reference standard are positive is $\alpha = \alpha_1 + \alpha_2$ (i.e., the sum of the proportion of positive index and reference test results when the condition is present (α_1) and absent (α_2)), and similarly for the other groups. A white colored box and the subscript 1 is used when the reference standard result matches the true status of the condition of interest; a grey colored box and the subscript 2 is used when it does not.

accuracy). More accurately, a study of diagnostic accuracy tries to estimate quantities that are functions of the eight probabilities.

Diagnostic Accuracy—the Case of the "Perfect" Reference Standard. A "perfect" reference standard would be infallible, always match the condition of interest, and, thus, in Figure 1 the proportions in the grey boxes (α_2, β_2, γ_2 and δ_2) would be zero. The data in the 2 × 2 table are then sufficient to estimate the four remaining probabilities (α_1, β_1, γ_1, and δ_1). Because the four probabilities necessarily sum to 1, it is sufficient to estimate any three. In practice, one estimates three other parameters, which are functions of the probabilities in the cells, namely, the sensitivity and specificity of the index test and the prevalence of the condition of interest (Table 1). If the counts in the table 2 × 2 are available (e.g., from a cohort study assessing the index test's performance), one can estimate the sensitivity and the specificity of the index test in a straightforward manner: $\widehat{Se_{index}} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$, and $\widehat{Se_{index}} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$, respectively.

Diagnostic Accuracy—the Case of the "Imperfect" Reference Standard. Only rarely are we sure that the reference standard is a perfect reflection of the truth. Most often in our assessments we accept some degree of misclassification by the reference standard, implicitly accepting it as being "as good as it gets". Table 2 lists some situations where we might question the validity of the reference standard. Unfortunately, there are no hard and fast rules for judging the adequacy of the reference standard; systematic reviewers should consult content experts in making such judgments.

Table 1. Parameterization When the Reference Standard is Assumed "Perfect" ("Gold Standard")

	Reference standard (+)	Reference standard (-)
Index test (+)	$\underbrace{p \times Se_{index}}_{\alpha_1} + \underbrace{0}_{\alpha_2}$	$\underbrace{(1-p) \times (1-Sp_{index})}_{\gamma_1} + \underbrace{0}_{\gamma_2}$
Index test (-)	$\underbrace{p \times (1-Se_{index})}_{\beta_1} + \underbrace{0}_{\beta_2}$	$\underbrace{(1-p) \times Sp_{index}}_{\delta_1} + \underbrace{0}_{\delta_2}$

Only three unknowns exist: the sensitivity and specificity of the index test (Se_{index} and Sp_{index}, respectively) and the disease prevalence (p). The under-braces refer to the probabilities of the 8 strata in Figure 1. These can be estimated from test results in a study, as discussed in the appendix of another paper in this supplement of the journal.[5]

Table 2. Situations Where One Can Question the Validity of the Reference Standard

Situation	Example
The reference method yields different measurements over time or across settings	Briefly consider the diagnosis of obstructive sleep apnea, which typically requires a high Apnea-Hypopnea Index (AHI, an objective measurement), and the presence of suggestive symptoms and signs. However, there is large night-to-night variability in the measured AHI, and there is also substantial variability between raters and between labs
The condition of interest is variably defined	This can be applicable to diseases that are defined in complex ways or qualitatively (e.g., based both on symptom intensity and on objective measurements). Such an example could be a complex disease such as psoriatic arthritis. There is no single symptom, sign, or measurement that suffices to make the diagnosis of the disease with certainty. Instead a set of criteria including symptoms, signs, imaging and laboratory measurements are used to identify it. Unavoidably, diagnostic criteria will be differentially applied across studies, and this is a potential explanation for the varying prevalence of the disease across geographic locations[34] and over time
The new method is an improved version of a usually applied test	Older methodologies for the measurement of parathyroid hormone (PTH) are being replaced by newer, more specific ones. PTH measurements with different methodologies do not agree very well.[35] Here, it would be wrong to assume that the older version of the test is the reference standard for distinguishing patients with high PTH from those without

Table 3 shows the relationship between the sensitivity and specificity of the index and reference tests and the prevalence of the condition of interest when the results of the index and reference tests are independent among those with and without the condition of interest ("conditional independence", one of several possibilities). For conditionally independent tests, estimates of sensitivity and specificity from the standard formulas ("naïve estimates") are always smaller than the true values (see an example in Fig. 2, and later for a more detailed discussion).

Options for Systematic Reviewers

So how should one approach the challenge of synthesizing information on diagnostic or prognostic tests when the purported "reference standard" is judged to be inadequate? At least four options exist. The first two change the framing of the problem, and forgo the classical paradigm for evaluating test performance. The third and fourth work within the classical paradigm and rely on qualifying the interpretation of results, or on mathematical adjustments:

1. Forgo the classical paradigm; assess the index test's ability to predict patient relevant outcomes instead of test accuracy (i.e., treat the index test as a predictive instrument).[5,6] This reframing applies when outcome information (usually on long term outcomes) exists, and the measured patient outcomes are themselves valid. If so, the approach to such a review is detailed in Chapter 11 in this supplement of the Journal.[7]
2. Forgo the classical paradigm; assess simply whether the results of the two tests (index and reference) agree or disagree (i.e., treat them as two alternative measurement methods). Instead of calculating sensitivity and specificity one would calculate statistics on test concordance, as mentioned later.

Table 3. Parameterization When the Reference Test is Assumed to be Imperfect, and the Index and Reference Test Results are Assumed Independent within the Strata of the Condition of Interest

	Reference test (+)	Reference test (-)
Index test (+)	$\underbrace{p \times Se_{ref} \times Se_{index}}_{\alpha_1} + \underbrace{(1-p) \times (1-Sp_{ref}) \times (1-Sp_{index})}_{\alpha_2}$	$\underbrace{(1-p) \times Sp_{ref} \times (1-Sp_{index})}_{\gamma_1} + \underbrace{p \times (1-Se_{ref}) \times Se_{index}}_{\gamma_2}$
Index test (-)	$\underbrace{p \times Se_{ref} \times (1-Se_{index})}_{\beta_1} + \underbrace{(1-p) \times (1-Sp_{ref}) \times Sp_{index}}_{\beta_2}$	$\underbrace{(1-p) \times Sp_{ref} \times Sp_{index}}_{\delta_1} + \underbrace{p \times (1-Se_{ref}) \times (1-Se_{index})}_{\delta_2}$

We now have five unknowns: the sensitivity and specificity of the index test (Se_{index} and Sp_{index}, respectively) and of the reference test (Se_{ref} and Sp_{ref}, respectively), and the disease prevalence (p). The under-braces refer to the probabilities of the 8 strata in Figure 1. Note that sensitivity and specificity always refer to the (unknown) true status of the condition of interest. Further, the results of the tests are assumed to be independent given the true status of the condition of interest. The cross-tabulation of the results of the index and reference tests is not sufficient to specify the problem, and additional information is necessary. It is easy to see that if the reference test is "perfect" ($Se_{ref} = 1$, $Sp_{ref} = 1$), one obtains the parameterization in Table 1. If the results of the index and reference tests are not independent among units with or without the condition of interest, the formulas in Table 1 change; in fact, several parameterizations are possible.[18,25–31]

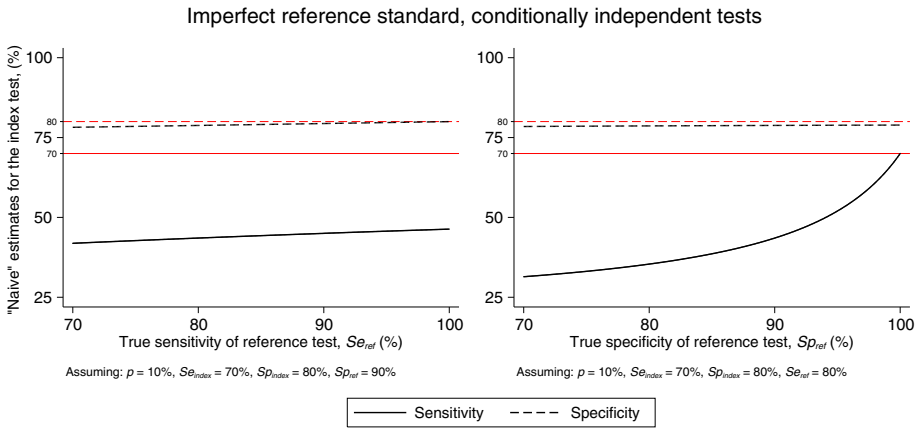

Figure 2. "Naïve" estimates versus true values for the performance of the index test with an imperfect reference standard. Se_{index} and Sp_{index}: sensitivity and specificity of the index test, respectively; Se_{ref} and Sp_{ref}: sensitivity and specificity of the reference test, respectively; p: disease prevalence. If the results of the index and reference tests are independent conditional on disease status, the "naïve" estimates for the performance of the index test are underestimates. The thin reference lines are the true sensitivity (solid) and specificity (dashed) of the index test. Note that the "naïve" estimate for the sensitivity and specificity of the index test approach the true values as the sensitivity and specificity of the reference test approaches 100%. In the left plot the "naïve" estimate of sensitivity does not reach 70% (the true value) when the sensitivity of the reference test, Se_{ref}, is 100%, because the specificity of the reference test is not perfect (Sp_{ref}=90%). Similarly, on the plot on the right, the specificity of the index test does not reach the true value of 80% when the specificity of the reference test, Sp_{ref}, is 100%, because the sensitivity of the reference test is not perfect (Se_{ref}=80%). The "naïve" estimates would be the same as the true values only if both the sensitivity and the specificity of the reference test are 100%.

3. Work within the classical paradigm, and calculate "naïve estimates" of the index test's sensitivity and specificity from each study, but qualify study findings.
4. Adjust the "naïve" estimates of sensitivity and specificity of the index test to account for the imperfect reference standard.

Our subjective assessment is that, when possible, the first option is preferred as it recasts the problem into one that is inherently clinically meaningful. The second option may be less clinically meaningful, but is a defensible alternative to treating an inadequate reference standard as if it were effectively perfect. The third option is potentially subject to substantial bias, which is especially difficult to interpret when the results of the test under review and the "reference standard" are not conditionally independent (i.e., an error in one is more or less likely when there is an error in the other). The forth option would be ideal if the adjustment methods were successful (i.e., eliminated biased estimates of sensitivity and specificity in the face of an imperfect reference standard). However, the techniques available necessarily require information that is typically not included in the reviewed studies, and require advanced statistical modeling.

1. Assess index test's ability to predict patient-relevant outcomes instead of test accuracy.

   This option is not universally possible. Instead of assessing the diagnostic or screening performance of the test, it quantifies the impact of patient management strategies that include testing on (usually long term) clinical outcomes. When it is possible and desirable to recast the evaluation question as an assessment of a tests ability to predict health outcomes, there are specific methods to consider when performing the assessment. For a more detailed discussion, the reader is referred to Paper 11 in this supplement of the Journal.[7]
2. Assess the concordance of difference tests instead of test accuracy

   Here, the index and reference tests are treated as two alternative measurement methods. One explores how well one test agrees with the other test(s), and perhaps if one test can be used in the place of the other. Assessing concordance may be the only meaningful option if none of the compared tests is an obvious choice for a reference standard (e.g., when both tests are alternative methodologies to measure the same quantity).

   In the case of categorical test results, one can summarize the extent of agreement between two tests using Cohen's κ statistic (a measure of categorical agreement which takes into account the probability that some agreement will occur by chance). A meta-analysis κ of statistics may also be considered to supplement a systematic review[8]; because it is not common practice in the medical literature, such a meta-analysis should be explained and interpreted in some detail.

   In the case of continuous test results, one is practically limited by the data available. If individual data points are available or extractable (e.g., in appendix tables or by digitizing plots) one can directly compare measurements with one test versus measurements with the other test. One way is to perform an appropriate regression to obtain an equation for translating the measurements with one test to the

measurements of the other. Because both measurements have random noise, an ordinary least squares regression is not appropriate; it treats the "predictor" as fixed and error-free, and thus underestimates the slope of the relationship between the two tests. Instead one should use a major axis or similar regression,[9–12] or more complex regressions that account for measurement error; consulting a statistician is probably wise. An alternative and well-known approach is to perform difference versus average analyses (Bland–Altman-type of analyses[13–15]. A qualitative synthesis of information from Bland–Altman plots can be quite informative (see example).[16] As of this writing the authors have not encountered any methods for incorporating difference versus average information from multiple studies.

If individual data points are not available, one has to summarize study-level information of the agreement of individual measurements. Of importance, care is needed when selecting which information to abstract. Summarizing results from major axis regressions or Bland–Altman analyses is probably informative. However, other metrics are not necessarily as informative. For example, Pearson's correlation coefficient, while often used to "compare" measurements with two alternative methods, is not a particularly good metric for two reasons: First, it does not inform on the slope of the line describing the relationship between the two measurements; it informs on the degree of linearity of the relationship. Further, its value can be high (e.g., >0.90) even when the differences between the two measurements are clinically important. Thus, one should be circumspect in using and interpreting a high Pearson's correlation coefficient for measurement comparisons.

3. Qualify the interpretation of "naïve" estimates of the index test's performance

   This option is straightforward. One could obtain "naïve" estimates of index test performance and make qualitative judgments on the direction of the bias of these "naïve" estimates.

Tests with Independent Results within the Strata of the Disease. We have seen already in Table 3 that, when the results of the index and reference test are independent among those with and without the disease (conditional independence), the "naïve" sensitivity and specificity of the index test is biased down. The "more imperfect" the reference standard, the greater the difference between the "naïve" estimates and true test performance for the index test (Fig. 2).

Tests with Correlated Results within the Strata of the Disease. When the two tests are correlated conditional on disease status, the "naïve" estimates of sensitivity and specificity can be overestimates or underestimates, and the formulas in Table 3 do not hold. They can be overestimates when the tests tend to agree more than expected by chance. They can be underestimates when the correlation is relatively small, or the tests disagree more than expected by chance.

A clinically relevant example is the use of prostate-specific antigen (PSA) to detect prostate cancer. PSA levels have been used to detect the presence of prostate cancer, and over the years, a number of different PSA detection methods have been developed. However, PSA levels are not elevated in as many as 15 % of individuals with prostate cancer, making PSA testing prone to misclassification error.[17] One explanation for these misclassifications (false-negative results) is that obesity can reduce serum PSA levels. The cause of misclassification (obesity) will likely affect all PSA detection methods—patients who do not have elevated PSA by a new detection method are also likely to not have elevated PSA by the older test. This "conditional dependence" will likely result in an overestimation of the diagnostic accuracy of the newer (index) test. In contrast, if the newer PSA detection method was compared to a non-PSA based reference standard that would not be prone to error due to obesity, such as prostate biopsy, conditional dependence would not be expected and estimates of diagnostic accuracy of the newer PSA method would likely be underestimated if misclassification occurs.

Because of the above, researchers should not assume conditional independence of test results without justification, particularly when the tests are based upon a common mechanism (e.g., both tests are based upon a particular chemical reaction, so that something which interferes with the reaction for one of the tests will likely interfere with the other test as well).[18]

4. Adjust or correct the "naïve" estimates of sensitivity and specificity

   Finally, one can mathematically adjust or correct the "naïve" estimates of sensitivity and specificity of the index test to account for the imperfect reference standard. The 2 × 2 cross-tabulation of test results is not sufficient to estimate the true sensitivities and specificities of the two tests, the prevalence of the conditions of interest, and correlations between sensitivities and specificities among those with and without the condition of interest. Therefore, additional information is needed. Several options have been explored in the literature. The following is by no means a comprehensive description; it is just an outline of several of the numerous approaches that have been proposed.

The problem is much easier if one can assume conditional independence for the results of the two tests, and further, that some of the parameters are known from prior knowledge. For example, one could assume that the sensitivity and specificity of the reference standard to detect true disease status is known from external sources, such as other studies,[19] or that the specificities for both tests are known (from prior studies) but the sensitivities are unknown.[20] In the same vein one can encode knowledge from

external sources with *prior distributions instead of fixed values*, using Bayesian inference.[21–24] Using a whole distribution of values rather than a single fixed value is less restrictive and probably less arbitrary. The resulting posterior distribution provides information on the specificities and sensitivities of both the index test and the reference standard, and of the prevalence of people with disease in each study.

When conditional independence cannot be assumed, the conditional correlations have to be estimated as well. Many alternative parameterizations for the problem have been proposed.[18,25–31] It is beyond the scope of the paper to describe them. Again, it is advisable to seek expert statistical help when considering such quantitative analyses, as modeling assumptions can have unanticipated implications[32] and model misspecification can result in biased estimates.[33]

Illustration

As an illustration we use a systematic review on the diagnosis of obstructive sleep apnea (OSA) in the home setting.[16] Briefly, OSA is characterized by sleep disturbances secondary to upper airway obstruction. It is prevalent in 2 to 4 % of middle-aged adults, and has been associated with daytime somnolence, cardiovascular morbidity, diabetes and other metabolic abnormalities, and increased likelihood of accidents and other adverse outcomes. Treatment (e.g., with continuous positive airway pressure) reduces symptoms, and, hopefully, long term risk for cardiovascular and other events. There is no "perfect" reference standard for OSA. The diagnosis of OSA is typically established based on suggestive signs (e.g. snoring, thick neck) and symptoms (e.g., somnolence), and in conjunction with an objective assessment of breathing patterns during sleep. The latter is by means of facility-based polysomnography, a comprehensive neurophysiologic study of sleep in the lab setting. Most commonly, polysomnography quantifies one's apnea-hypopnea index (AHI) (i.e., how many episodes of apnea [no airflow] or hypopnea [reduced airflow] a person experiences during sleep). Large AHI is suggestive of OSA. At the same time, portable monitors can be used to measure AHI instead of facility based polysomnography.

Identifying (Defining) the Reference Standard. One consideration is what reference standard is most common, or otherwise "acceptable", for the main analysis. In all studies included in the systematic review, patients were enrolled only if they had suggestive symptoms and signs (although it is likely that these were differentially ascertained across studies). Therefore, in these studies, the definition of "sleep apnea" is practically equivalent to whether people have a "high enough" AHI.

Most studies and some guidelines define AHI≥15 events per hour of sleep as suggestive of the disease, and this is the cut-off selected for the main analyses. In addition, identified studies used a wide range of cut-offs in the reference method to define sleep apnea (including 5, 10, 15, 20, 30, and 40 events per hour of sleep). As a sensitivity analysis, the reviewers decided to summarize studies also according to the 10 and the 20 events per hour of sleep cut-offs; the other cut-offs were excluded because data was sparse. It is worth noting that, in this case, the exploration of the alternative cut-offs did not affect the results or conclusions of the systematic review, but did require substantial time and effort.

Deciding How to Summarize the Findings of Individual Studies and How to Present Findings. The reviewers calculated "naïve" estimates of sensitivity and specificity of portable monitors, and qualified their interpretation (option 3). They also performed complementary analyses outside the classical paradigm for evaluating test performance to describe the concordance of measurements with portable monitors ("index" test) and facility-based polysomnography ("reference" test; this is option 2).

Qualitative Analyses of "Naïve" Sensitivity and Specificity Estimates. The reviewers depicted graphs of the "naïve" estimates of sensitivity and specificity in the ROC space (see Fig. 3). These graphs suggest a high "sensitivity" and "specificity" of portable monitors to diagnose AHI≥15 events per hour with facility-based polysomnography.

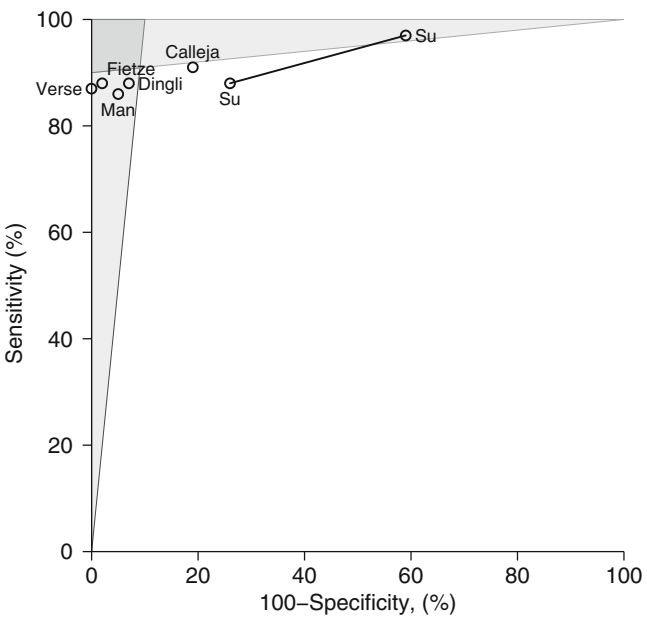

Figure 3. "Naïve" estimates of the ability of portable monitors versus laboratory-based polysomnography to detect AHI > 15 events/hour. These data are on a subset of studies from the systematic review used in the illustration (Studies that used manual scoring or combined manual and automated scoring for a type III portable monitor). "Naive" sensitivity/specificity pairs from the same study (obtained with different cut-offs for the portable monitor) are connected with lines. Studies lying on the left lightly shaded area have a positive likelihood ratio of 10 or more. Studies lying on the top lightly shaded area have a negative likelihood ratio of 0.1 or less. Studies lying on the intersection of the grey areas (darker grey polygon) have both a positive likelihood ratio more than 10 and a negative likelihood ratio less than 0.1.

However, it is very difficult to interpret these high values. First, there is considerable night-to-night variability in the measured AHI, as well as substantial between-rater and between-lab variability. Second, it is not easy to deduce whether the "naïve" estimates of "sensitivity" and "specificity" are underestimates or overestimates compared to the unknown "true" sensitivity and specificity to identify "sleep apnea."

The systematic reviewers suggested that a better answer would be obtained by studies that perform a clinical validation of portable monitors (i.e., their ability to predict patients' history, risk propensity, or clinical profile—this would be option 1) and identified this as a gap in the pertinent literature.

Qualitative Assessment of the Concordance Between Measurement Methods. The systematic reviewers decided to summarize Bland–Altman type analyses to obtain information on whether facility-based polysomnography and portable monitors agree well enough to be used interchangeably. For studies that did not report Bland–Altman plots, the systematic reviewers performed these analyses using patient-level data from each study, extracted by digitizing plots. An example is shown in Figure 4. The graph plots the differences between the two measurements against their average (which is the best estimate of the true unobserved value). An important piece of information from such analyses is the range of values defined by the 95 percent limits of agreement (i.e., the region in which 95 % of the differences are expected to fall). When the 95 % limits of agreement are very broad, the agreement is suboptimal (Fig. 4).

Figure 5 summarizes such plots across several studies. For each study, it shows the mean difference in the two measurements (mean bias) and the 95 % limits of agreement. The qualitative conclusion is that the 95 % limits of agreement are very wide in most studies, suggesting great variability in the measurements with the two methods.

Thus, AHI measurements with the two methods generally agree on who has 15 or more events per hour of sleep (which is a low AHI). They disagree on the exact measurement among people who have larger measurements on average: One method may calculate 20 and the other 50

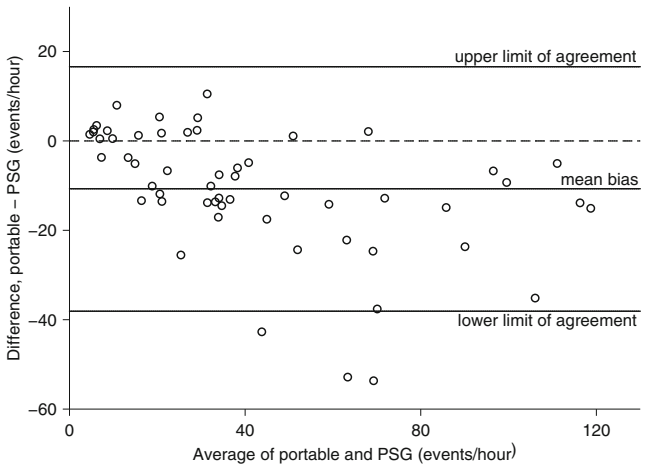

Figure 4. Illustrative example of a difference versus average analysis of measurements with facility-based polysomnography and portable monitors. Digitized data from an actual study where portable monitors (Pro-Tech PTAF2 and Compumedics P2) were compared with facility-based polysomnography (PSG).[16] The dashed line at zero difference is the line of perfect agreement. The mean bias stands for the average systematic difference between the two measurements. The 95 % limits of agreement are the boundaries within which 95 % of the differences lie. If these are very wide and encompass clinically important differences, one may concur that the agreement between the measurements is suboptimal. Note that the spread of the differences increases for higher measurement values. This indicates that the mean bias and 95 % limits of agreement do not describe adequately the differences between the two measurements; differences are smaller for smaller values and larger for larger AHI values. In this example mean bias = -11 events/hour (95 % limits of agreement: –38, 17), with statistically significant dependence of difference on average (Bradley-Blackwood F test, $p<0.01$).

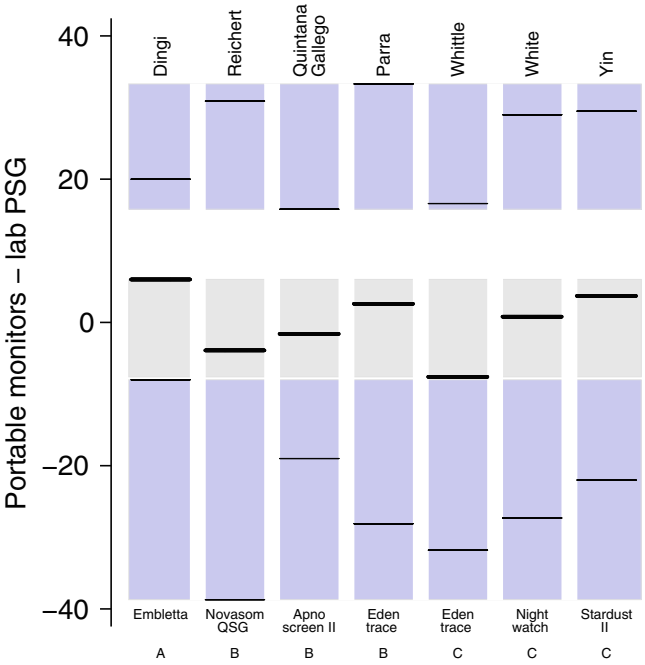

Figure 5. Schematic representation of the mean bias and limits of agreement across several studies. Schematic representation of the agreement between portable monitors and facility-based polysomnography as conveyed by difference versus average analyses across seven studies (the study of Fig. 4 is not included). The study author and the make of the monitor are depicted in the upper and lower part of the graph, respectively. The difference versus average analyses from each study are represented by three horizontal lines: a thicker middle line (denoting the mean bias); and two thinner lines, which represent the 95 % limits of agreement and are symmetrically positioned above and below the mean bias line. The figure facilitates comparisons of the mean bias and the 95 % limits of agreement across the studies by means of colored horizontal zones. The middle light-gray-colored zone shows the range of the mean bias in the seven studies, which is from +6 events per hour of sleep in the study by Dingi et al. (Embletta monitor) to -8 events per hour of sleep in the study by Whittle et al. (Edentrace monitor). The uppermost and lowermost shaded areas show the corresponding range of the upper 95 % limits of agreement (upper shaded zone) and the lower 95 % limits of agreement (lower shaded zone) in the seven studies.

events per hour of sleep for the same person. The two methods are expected to disagree on who has AHI for those with >20, >30, or >40 events per hour.

SUMMARY

In approaching a systematic review of the performance of a medical test, one is often faced with a reference standard which itself is subject to error. Four potential approaches are suggested:

1. If possible, recast the assessment task in which the index test is used as an instrument to predict clinical outcomes. This reframing is potentially applicable only when measured patient outcomes are themselves valid. If so, the approach to such a review is detailed in Paper 11 in this supplement of the journal.
2. Assess the condordance in the results of the index and reference tests (i.e., treat them as two alternative measurement methods).
3. Calculate "naïve estimates" of the index test's sensitivity and specificity from each study, but qualify study findings.
4. Adjust the "naïve" estimates of sensitivity and specificity of the index test to account for the imperfect reference standard.

Systematic reviewers should decide which of the four options is more suitable for evaluating the performance of an index test versus an "imperfect" reference standard. To this end, the following considerations should be taken into account in the planning stages of the review: First, it is possible that multiple (imperfect) reference standard tests, or multiple cutoffs for the same reference test, are available. If an optimal choice is not obvious, the systematic reviewer should consider assessing more than one reference standard, or more than one cutoff for the reference test (as separate analyses). Whatever the choice, the implications of using the reference standard(s) should be described explicitly. Second, the reviewers should decide which option(s) for synthesizing test performance is (are) appropriate. The four options need not be mutually exclusive, and in some cases can be complementary (e.g., a "naïve" and "adjusted" analyses would reinforce assessments of a test if they both lead to similar clinical implications.) Finally, most of the analyses alluded to in option 4 would require expert statistical help; further, we have virtually no empirical data on the merits and pitfalls of methods that mathematically adjust for an "imperfect" reference standard. In our opinion, in most cases options 1-3 would provide an informative summary of the data.

Acknowledgment: *This manuscript is based on work funded by the Agency for Healthcare Research and Quality (AHRQ). Both authors are members of AHRQ-funded Evidence-based Practice Centers. The opinions expressed are those of the authors and do not reflect the official position of AHRQ or the U.S. Department of Health and Human Services.*

Conflict of Interest: *The authors declare that they have no conflict of interest.*

Corresponding Author: *Thomas A. Trikalinos, MD; Center for Evidence-based Medicine, and Department of Health Service, Policy and Practice, Brown University, G-S121-7, Providence, RI 02912, USA (e-mail: thomas_trikalinos@brown.edu).*

REFERENCES

1. **Bossuyt PM.** Interpreting diagnostic test accuracy studies. Semin Hematol. 2008;45(3):189–195.
2. **Bossuyt PM, Reitsma JB, Bruns DE, et al.** Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Radiology. 2003;226(1):24–28.
3. **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(50):iii, ix-51.
4. **Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J.** Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3):189–202.
5. **Trikalinos TA, Balion CM, Coleman CI, et al.** Chapter 8: Meta-analysis of test performance when there is a "Gold Standard." J Gen Internal Med. 2012; doi: 10.1007/s11606-012-2029-1
6. **Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM.** A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009;62(8):797–806.
7. **Jonas DE, Wilt TJ, Taylor BC, Wilkins TM, Matchar DB.** Chapter 11: Challenges in and principles for conducting systematic reviews of genetic tests used as predictive indicators. J Gen Internal Med. 2011; doi: 10.1007/s11606-011-1898-z
8. **Sun S.** Meta-analysis of Cohen's kappa. Health Serv Outcomes Res Method. 2011;11:145–163.
9. **Sokal RR, Rohlf EF.** Biometry. New York: Freeman; 1981.
10. **Bablok W, Passing H, Bender R, Schneider B.** A general regression procedure for method transformation. Application of linear regression procedures for method comparison studies in clinical chemistry, Part III. J Clin Chem Clin Biochem. 1988;26(11):783–790.
11. **Linnet K.** Estimation of the linear relationship between the measurements of two methods with proportional errors. Stat Med. 1990;9(12):1463–1473.
12. **Linnet K.** Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. Clin Chem. 1998;44(5):1024–1031.
13. **Altman DG, Bland JM.** Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485.
14. **Bland JM, Altman DG.** Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8(2):135–160.
15. **Bland JM, Altman DG.** Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol. 2003;22(1):85–93.
16. **Trikalinos TA, Ip S, Raman G, Cepeda MS, Balk EM, D'Ambrosio C, et al.** Home diagnosis of obstructive sleep apnea-hypopnea syndrome. Evidence Report/Technology Assessment. Rockville, MD: Agency for Healthcare Research and Quality; 2007:1–127. Evidence Report/Technology Assessment. Ref Type: Report.
17. **Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL, et al.** Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. N Engl J Med. 2004;350(22):2239–2246.
18. **Vacek PM.** The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics. 1985;41(4):959–968.
19. **Gart JJ, Buck AA.** Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am J Epidemiol. 1966;83(3):593–602.
20. **Goldberg JD, Wittes JT.** The estimation of false negatives in medical screening. Biometrics. 1978;34(1):77–86.
21. **Gyorkos TW, Genta RM, Viens P, MacLean JD.** Seroepidemiology of Strongyloides infection in the Southeast Asian refugee population in Canada. Am J Epidemiol. 1990;132(2):257–264.

22. **Joseph L, Gyorkos TW.** Inferences for likelihood ratios in the absence of a "gold standard". Med Decis Making. 1996;16(4):412–417.
23. **Walter SD, Irwig L, Glasziou PP.** Meta-analysis of diagnostic tests with imperfect reference standards. J Clin Epidemiol. 1999;52(10):943–951.
24. **Walter SD, Irwig LM.** Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol. 1988;41(9):923–937.
25. **Dendukuri N, Joseph L.** Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics. 2001;57(1):158–167.
26. **Black MA, Craig BA.** Estimating disease prevalence in the absence of a gold standard. Stat Med. 2002;21(18):2653–2669.
27. **Dendukuri N, Hadgu A, Wang L.** Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Stat Med. 2009;28(3):441–461.
28. **Garrett ES, Eaton WW, Zeger S.** Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. Stat Med. 2002;21(9):1289–1307.
29. **Hui SL, Zhou XH.** Evaluation of diagnostic tests without gold standards. Stat Methods Med Res. 1998;7(4):354–370.
30. **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.
31. **Torrance-Rynard VL, Walter SD.** Effects of dependent errors in the assessment of diagnostic test performance. Stat Med. 1997;16(19):2157–2175.
32. **Toft N, Jorgensen E, Hojsgaard S.** Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Prev Vet Med. 2005;68(1):19–33.
33. **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(2):427–435.
34. **Alamanos Y, Voulgari PV, Drosos AA.** Incidence and prevalence of psoriatic arthritis: a systematic review. J Rheumatol. 2008;35(7):1354–1358.
35. **Cantor T, Yang Z, Caraiani N, Ilamathi E.** Lack of comparability of intact parathyroid hormone measurements among commercial assays for end-stage renal disease patients: implication for treatment decisions. Clin Chem. 2006;52(9):1771–1776.