Methods Guide for Medical Test reviews

Paper 9

Meta-analysis of Test Performance Evidence When There is an Imperfect Reference Standard

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

DRAFT

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

Effectiveness and Comparative Effectiveness Reviews, systematic reviews of existing research on the effectiveness, comparative effectiveness, and comparative harms of different medical tests, are intended to provide relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. In an effort to improve the transparency, consistency, and scientific rigor of the work of the Effective Health Care (EHC) Program, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and highlights particular areas where the inherently different qualities of medical tests necessitate a different or variation of the approach to systematic review compared to a review on treatments. We provide guidance in stepwise fashion for those conducting a systematic review.

The *Medical Test Methods Guide* is a living document, and will be updated as further empirical evidence develops and our understanding of better methods improves. Comments and suggestions on the *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

DRAFT

Paper 9. Meta-analysis of Test Performance Evidence When There is an Imperfect Reference Standard

To evaluate the performance of a medical test (index test), we have to compare its results with the "true" status of every tested individual or specimen. Sometimes, this true status is directly observable; for example, when we use a test to predict short-term mortality after a procedure, or when we verify whether a suspect lesion is malignant with an excisional biopsy. However, in many cases the true status of the tested unit is judged based on another test as a reference method. Problems can arise when the reference standard test does not mirror truth adequately well: in such a case, we will be measuring the performance of the index test against a faulty standard, and we are bound to err. In fact, the further the reference standard test deviates from the truth, the poorer our estimate of the index test's performance will be. This is otherwise known as "reference standard bias."[1-4]

Common Challenges

Only rarely are we absolutely sure that a reference standard test is a perfect reflection of the truth. Most often, we are very comfortable with overlooking small or moderate misclassifications by the reference standard. In fact, this is exactly what we do, implicitly, when we calculate the index test's sensitivity, specificity, and related quantities. But how should we approach the evaluation of a diagnostic or prognostic test when the reference standard itself performs (too) poorly? Table 9-1 lists some common situations where we might question the validity of the reference standard. We do not discuss the case of a "missing gold standard," where the reference standard is guided by the results of the index test and not universally applied (also known as verification bias).

Table 9-1. Situations where the validity of the reference standard is in question

Situation	Example
The reference standard test 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 between-rater and between-laboratory variability.
The definition of the "disease" being tested for is idiosyncratic to some extent.	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, and imaging and laboratory measurements are used to identify it. Unavoidably, diagnostic criteria will be differentially applied across studies, and this is a potential explanation of the varying prevalence of the disease across geographic locations.[5]

DRAFT

Situation	Example
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.[6] In this case, 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.

Principles for Addressing the Challenges

Principle 1: Favor the Simplest Analysis That Properly Summarizes the Data

There are several ways to approach a systematic review of medical tests when the reference standard is a poor approximation of the truth, or when no test can be regarded as a reference standard. One can choose among the following options depending on the topic at hand:

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

(2) Adjust or correct the "naïve" estimates of sensitivity and specificity of the index test to account for the imperfect reference standard.

(3) Assess the concordance of difference tests instead of test accuracy.

Assess the test's ability to predict patient-relevant outcomes instead of the test's accuracy. Recast the review so that it does not aim to calculate estimates of test performance[a] but rather aims to assess whether test results predict relevant clinical data, such as history, future clinical events, and response to therapy.[2] Essentially, this implies ignoring all information that compares the index test with the imperfect reference standard. The rationale is that this information is not informative or interpretable. We will not discuss this option further. Paper 12 of the *Medical Test Guide* elaborates on the evaluation of prognostic tests.

Adjust or correct the naïve estimates of sensitivity and specificity. One can adjust or correct the naïve estimates of sensitivity and specificity of the index test to account for the imperfect reference standard. This implies that, apart from the index test, there is a test that can be regarded as a reference standard, albeit an imperfect one. In such case, one cannot calculate sensitivity (ability to maximize true positives) and specificity (ability to minimize false positives) of the test using only the observed data.

To overcome this problem, one must either use constraints on a subset of the parameters (i.e., assume that the sensitivity and specificity of, e.g., the reference standard to detect true disease status is known,[7] or that both specificities are known but the sensitivities are unknown[8]), or use an approach that treats the missing gold standard as latent data and estimate it by combining external information from other sources (prior distributions) with the available data using Bayesian inference.[9-11] These prior distributions provide a different type of constraint, which

[a] Test performance as meant in Paper 8: "accuracy" measures such as sensitivity and specificity of the index test or other metrics that can be derived from these quantities.

makes use of prior knowledge of the parameters and thus may be less arbitrary. The resulting posterior distribution provides information on the specificities and sensitivities of both the index test and the reference standard, and on the prevalence of people with the diseases in each study.

In addition, statistical adjustment requires meta-analysts either to assume "conditional independence" between the index and reference standard results (which often does not hold), or to supply data regarding the between-test correlation if conditional independence is not assumed. However, such correlation data are not commonly available.[3]

Assess the concordance of difference tests instead of test accuracy. Here, the reviewer is no longer interested in the sensitivity and specificity of the examined test but rather in how well it agrees with the other test(s) and, perhaps, whether one test can be used in place of the other. Assessing concordance may be the only realistic option if none of the compared tests is an obvious choice for a reference standard; for example, when both tests are alternative methodologies to measure the same quantity.

Depending on the topic at hand, one can summarize the extent of agreement between two tests using Cohen's kappa statistics (a measure of categorical agreement which takes into account the probability that some agreement will occur by chance), Bland-Altman plots (agreement of continuous measurements),[12-14] and interclass correlation (ICC) statistics. It is anticipated that, in many cases, quantitative summaries of concordance data will not be particularly meaningful or even possible; then, the review can be limited to a qualitative descriptive analysis of the diagnostic research available.

Principle 2: Qualify Findings To Avoid Misinterpretations

As mentioned above, when the reference standard is (grossly) misclassifying the "true" disease status, the usually calculated estimates of sensitivity and specificity (naïve estimates) are biased. The direction of the bias can be either upward or downward, and the magnitude will depend on the frequency of reference standard errors and the degree of correlation in errors between the index test and reference standard.

Conditionally independent tests. In the simplest case, the index test and the reference standard are independent conditional on disease status. In other words, the two tests do not tend to agree more (or less) than expected among people with the disease or people without the disease. Then the naïve estimates of sensitivity and specificity are underestimates.

Conditionally dependent tests. When the two tests are correlated conditional on disease status, the naïve estimates of sensitivity and specificity can be overestimates or underestimates. Overestimates can happen when the tests tend to agree more than expected by chance. Underestimates can happen when the correlation is relatively small or the tests disagree more than expected by chance.[b] It is common for researchers to assume conditional independence of the results of different tests. However, this assumption cannot be uniformly justified, particularly when the tests are based on a common mechanism (e.g., both tests are based on a particular

[b] We are hard pressed to find a good clinical example where tests disagree more than expected by chance; however, it is a theoretical possibility.

chemical reaction so that something that interferes with the reaction for one of the tests will likely interfere with the other as well).[15]

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 percent of individuals with prostate cancer, making PSA testing prone to misclassification error.[16] One explanation for these misclassifications (false-negative results) is that obesity can reduce serum PSA levels. In this situation, 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 were compared to a non-PSA-based reference standard that was not prone to error due to obesity, such as a 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.

Therefore, in deciding how to approach the evaluation of test performance in the face of an imperfect reference standard, consider the following guidance:

(1) If not established at the outset that one test constitutes the reference standard and if there are multiple alternatives for a reference standard, decide which reference standard is most common (or otherwise acceptable) for the main analysis.[c] Consider exploring alternative reference standards for completeness.
(2) Decide which is more informative for the user of the systematic review:
    a. To have estimates of sensitivity and specificity for the index test?
    b. To learn whether the index test and the reference test are concordant or interchangeable?
    c. Both?
(3) Keep analyses and presentations simple. There is a relative paucity of empirical data on the merits and pitfalls of using the more complex methods that account for imperfect reference standards. We recommend that, in the majority of cases, reviewers should perform at most simple analyses, followed by clear discussion and illustration using relevant examples. We believe that it is transparent and instructive to:
    a. Present naïve estimates of sensitivity, specificity, and related analyses.
    b. Discuss the expected direction and magnitude of the bias of the naïve estimates.

Illustration

In this section, we expand on the sleep apnea technology assessment report to illustrate the issue of an imperfect reference standard.[17] This is a useful case as there is no accepted reference standard for the diagnosis of sleep apnea. In this example, we will indicate how to apply the principles discussed above to the estimation of performance of a proposed test for this condition.

[c] This should include consideration of cut points for continuous or ordinal test results, or clusters for tests with multiple unordered categories.

Sleep apnea technology assessment. As noted above, one consideration is what reference standard is most common or otherwise acceptable for the main analysis. In this case, the diagnosis of sleep apnea is usually (but imperfectly) diagnosed when the patient has a high Apnea-Hypopnea Index (AHI, an objective measurement), together with suggestive symptoms and signs. In all reviewed studies, patients were enrolled only if they had suggestive symptoms and signs, although it is likely that these were differentially ascertained across studies. Therefore, the definition of sleep apnea is reduced to whether people have a high enough AHI.

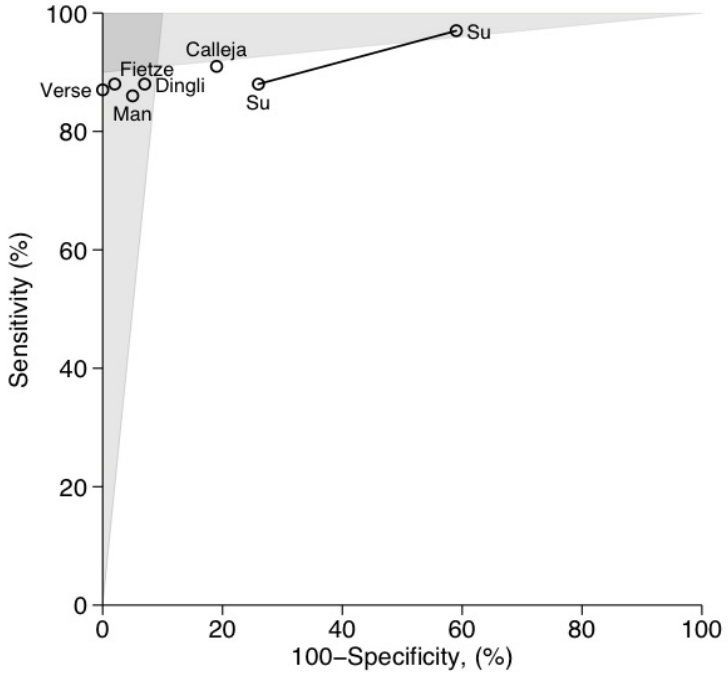
Defining the reference standard. Most studies and some guidelines define AHI ≥ 15 events per hour of sleep as being suggestive of the disease, and this is the cutoff selected for the main analyses. In addition, identified studies used a wide range of cutoffs in the reference method to define sleep apnea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

Deciding how to summarize and present the findings of individual studies. Following the aforementioned principles, the reviewers decided to calculate and interpret naïve estimates of sensitivity and specificity of portable monitors and to describe the concordance of measurements with portable monitors (index test) and facility-based polysomnography (reference test).

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 Figure 9-1). These graphs suggest a high "sensitivity" and "specificity" of portable monitors to diagnose AHI ≥ 15 events per hour with facility-based polysomnography. 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-laboratory 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 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), and they identified this as a gap in the pertinent literature.

DRAFT

Figure 9-1. Diagnostic ability of type III monitors in specialized sleep units to identify AHI>15 events/hour in laboratory-based polysomnography: manual or combined manual and automated scoring

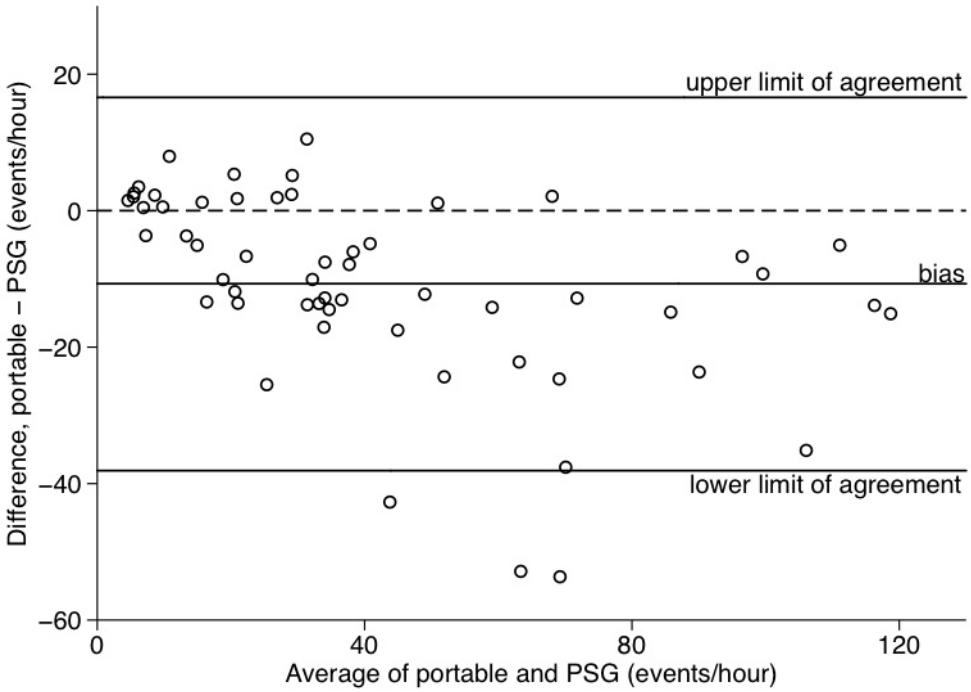


Sensitivity/specificity pairs from the same study (obtained with different cutoffs for the type III monitor) are connected with lines. These lines are not representative of the ROC curve of the pertinent studies. Studies lying on the left shaded area have a positive likelihood ratio (LR+) of 10 or more. Studies lying on the top shaded are have a negative likelihood ratio (LR-) of 0.1 or less. Studies lying on the intersection of the gray areas (darker gray polygon) have both LR+ > 10 and LR- < 0.1. The figure depicts studies that used manual scoring or combined manual and automated scoring for the type III monitor and a cutoff of 15 events/h as suggestive of sleep apnea in facility-based polysomnography.

Qualitative assessment of the concordance between measurement methods. Now, let us exemplify the alternative approach of summarizing analyses of agreement between the two methods. This approach essentially asks whether the two methods could be used interchangeably.

An obvious option is to record and summarize difference versus average analyses (Bland-Altman analyses) from the published studies. A Bland-Altman plot shows the differences between the two measurements against their average (which is the best estimate of the true unobserved value). An important concept in such analyses is the 95 percent limits of agreement. The 95 percent limits of agreement define the region in which 95 percent of the differences are expected to fall. When the 95 percent limits of agreement are very broad, the agreement is suboptimal (Figure 9-2).

Figure 9-2. Example of a difference versus average analysis of measurements with facility-based polysomnography and type IV monitors

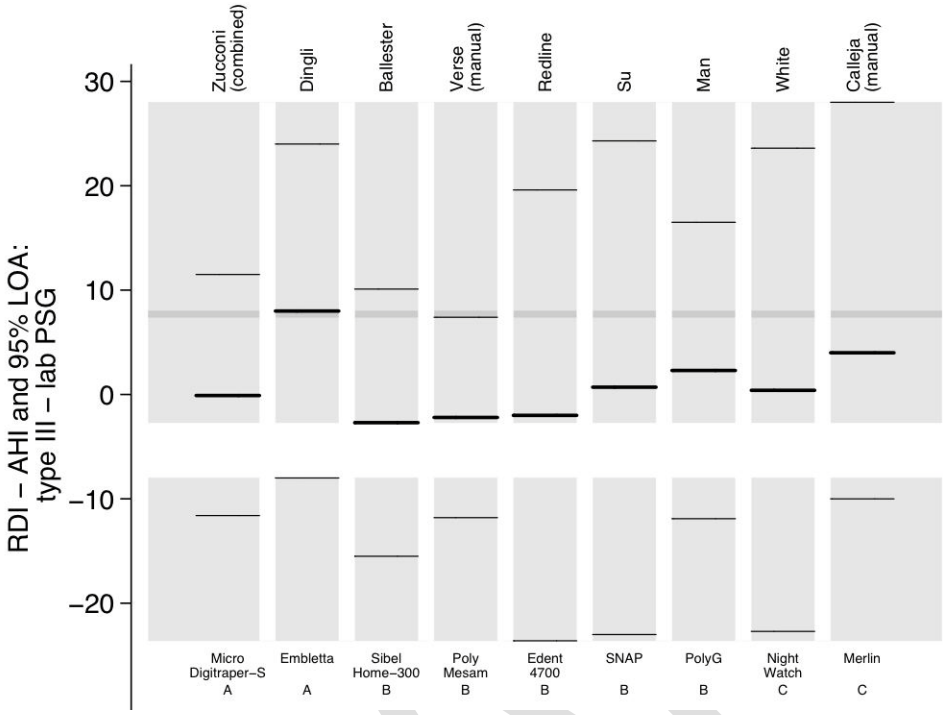


Digitized data from an actual study where type IV monitors (Pro-Tech® PTAF2 and Compumedics® P2) were compared with facility-based PSG.[17] The dashed line at zero difference is the line of perfect agreement. The mean bias stands for the average systematic difference between the two measurements. The 95 percent limits of agreement stand for the boundaries within which 95 percent 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 percent limits of agreement do not describe adequately the differences between the two measurements; differences are smaller for smaller AHI or RDI levels and larger for larger AHI or RDI levels. In this example, bias = -11 events/hour (95 percent limits of agreement: -38, 17), with statistically significant dependence of difference on average (Bradley-Blackwood F test, $p < 0.01$).

Figure 9-3 summarizes such plots across several studies. For each study, it shows the mean difference in the two measurements and the 95 percent limits of agreement. The qualitative conclusion is that the 95 percent limits of agreement are very wide in most studies, suggesting great variability in the measurements with the two methods.

DRAFT

Figure 9-3. Schematic representation of the mean bias and limits of agreement between facility-based polysomnography and type III monitors in specialized sleep units: studies that used manual scoring for the portable monitor



Schematic representation of the agreement between portable monitors and facility-based polysomnography as conveyed by difference versus average analyses. Each study is represented by three lines; these stand for the mean bias and the 95 percent limits of agreement from the difference versus average analyses. The upper, middle, and lower gray areas group the upper 95 percent limits of agreement, the mean difference, and the lower 95 percent limits of agreement, respectively. Note that the upper and middle gray areas overlap slightly. The make of the monitor and the overall study quality are also depicted in the lower part of the graph. Only studies that used both apneas and hypopneas in the definition of respiratory events for both monitors are shown.

This is at odds with the findings of the previous analysis, which suggested the portable monitors are good in identifying people with sleep apnea. The explanation may be that the two measurements generally agree on who has 15 or more events per hour of sleep (which is a low number). They disagree on the exact measurement among people who have larger measurements on average: One method may calculate 20 events and the other 50 events per hour of sleep for the same person. Such differences can become important when the definition of sleep apnea is at a higher level (20, 30, or 40 events per hour). This means that the Bland-Altman analyses are not particularly helpful when considering the cutoff of 15 events per hour for defining sleep apnea.

Summary

Key points are:

- When dealing with the case of an imperfect reference standard, it may be preferable to use a simple description of study results.

- Because of the challenges in the interpretation of the naïve estimates for sensitivity and specificity, a typical meta-analysis may not be the best way to summarize the findings.
- There is a relative paucity of empirical data on the merits and pitfalls of statistical methods that try to adjust for the imperfect reference standard. Therefore, it does not seem prudent to recommend that EPCs implement and adopt the complicated methods for the analysis of this case.

References

1. Bossuyt PM. Interpreting diagnostic test accuracy studies. Semin Hematol 2008;45(3):189-95.

2. Reitsma JB, Rutjes AW, Khan KS, et al. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol 2009;62(8):797-806.

3. Rutjes AW, Reitsma JB, Coomarasamy A, et al. 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, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140(3):189-202.

5. Alamanos Y, Voulgari PV, Drosos AA. Incidence and prevalence of psoriatic arthritis: a systematic review. J Rheumatol 2008;35(7):1354-8.

6. Cantor T, Yang Z, Caraiani N, et al. 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):1771-6.

7. 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.

8. Goldberg JD, Wittes JT. The estimation of false negatives in medical screening. Biometrics 1978;34(1):77-86.

9. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 2001;57(1):158-67.

10. Gyorkos TW, Genta RM, Viens P, et al. Seroepidemiology of Strongyloides infection in the Southeast Asian refugee population in Canada. Am J Epidemiol 1990;132(2):257-64.

11. Joseph L, Gyorkos TW. Inferences for likelihood ratios in the absence of a "gold standard". Med Decis Making 1996;16(4):412-7.

DRAFT

12. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311(7003):485.

13. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8(2):135-60.

14. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 2003;22(1):85-93.

15. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics 1985;41(4):959-68.

16. Thompson IM, Pauler DK, Goodman PJ, 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):2239-46.

17. Trikalinos TA, Ip S, Raman G, et al. Home diagnosis of obstructive sleep apnea-hypopnea syndrome. Technology Assessment. Rockville, MD: Agency for Healthcare Research and Quality. August 2007. Available at: http://www.cms.hhs.gov/determinationprocess/downloads/id48TA.pdf. Accessed July 9, 2010.

DRAFT