Medical Tests—White Paper Series

Decision-Analytic Modeling to Evaluate Benefits and Harms of Medical Tests—Uses and Limitations

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of the Agency for Healthcare Research and Quality (AHRQ). Therefore, no statement in this report should be construed as an official position of AHRQ or the U.S. Department of Health and Human Services.

This report has been published in edited form: Trikalinos TA, Siebert U, Lau J. Decision-analytic modelling to evaluate benefits and harms of medical tests: uses and limitations. Med Decis Making 2009 Sep-Oct;29(5):E22-E29. Epub 2009 Sep 4.

Suggested citation: Trikalinos TA, Siebert U, Lau J. Decision-analytic modelling to evaluate benefits and harms of medical tests. Medical Tests—White Paper Series. Agency for Healthcare Research and Quality: Rockville, MD. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=350

Decision-Analytic Modeling to Evaluate Benefits and Harms of Medical Tests—Uses and Limitations

Authors:

Thomas A. Trikalinos, M.D.[a]
Uwe Siebert, M.D., M.P.H., M.Sc., Sc.D.[b,c]
Joseph Lau, M.D.[a]

[a]Tufts Evidence-based Practice Center and Center for Clinical Evidence Synthesis, Tufts Medical Center, Boston, MA

[b]Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School and Center for Health Decision Science, Department of Health Policy and Management, Harvard School of Public Health, Boston, MA

[c]Department of Public Health Medical Decision Making and Health Technology Assessment UMIT—University for Health Sciences, Medical Informatics and Technology, Hall I.T., Austria

Decision-Analytic Modeling to Evaluate Benefits and Harms of Medical Tests—Uses and Limitations

Abstract

The clinical utility of medical tests is measured by whether the information they provide affects patient-relevant outcomes. To a large extent, effects of medical tests are indirect in nature. In principle, a test result affects patient outcomes mainly by influencing treatment choices. This indirectness in the link between testing and its downstream effects poses practical challenges to comparing alternative test-and-treat strategies in clinical trials. Keeping in mind the broader audience of researchers who perform comparative effectiveness reviews and technology assessments, we summarize the rationale for and pitfalls of decision modeling in the comparative evaluation of medical tests by using specific examples. Modeling facilitates the interpretation of test performance measures by connecting the link between testing and patient outcomes, accounting for uncertainties and explicating assumptions, and allowing the systematic study of tradeoffs and uncertainty. We discuss challenges encountered when modeling test-and-treat strategies, including, but not limited to, scarcity of data on important parameters, transferring estimates of test performance across studies, choosing modeling outcomes, and obtaining summary estimates for test performance data.

Introduction

The value of any medical test is ultimately measured by whether the information it provides affects patient-relevant outcomes such as morbidity, mortality, or health-related quality of life. Although testing in itself can affect outcomes directly,[1] most of its impact is indirect. In principle, test results influence downstream clinical decisions that will eventually determine patient outcomes. From this point of view, test performance (as conveyed by sensitivity, specificity, positive and negative likelihood ratios, or other metrics) is only a surrogate endpoint. The link between test results and their induced downstream effects has to be supported, theoretically or empirically, on a case-by-case basis.

Arguably, the most robust empirical demonstration of the utility of a medical test is through a properly designed randomized trial[2-5] that compares patient management with the test vs. one or more alternative strategies. In practice such trials are not routinely performed, because they are often deemed unattainable.[3,5] Obstacles are posed by the indirectness of the link between testing and clinical outcomes and the plethora of alternative test-and-treat strategies that are reasonable to contrast.[6] Observational studies of patient management strategies are also uncommonly performed, and further, selection bias and confounding can threaten their internal validity and generalizability.

Limited by the existing literature, systematic reviews of medical tests summarize performance characteristics rather than effects on patient outcomes.[7] However, the link between test performance and patient-relevant outcomes is typically complex. High test performance does not guarantee that physicians will act according to test results, that patients will adhere to recommendations, or that the chosen interventions will be effective. Moreover, when comparing strategies that utilize alternative tests, differences in test performance do not necessarily translate to corresponding differences in patient-relevant outcomes.

For the majority of tests and clinical settings, the link between test performance and patient-relevant outcomes must be deduced from evidence reported in different studies. In addition, health care decisions have to be made irrespective of evidence availability or unavailability and have to account for many factors beyond test performance and treatment effectiveness. Transparent and reproducible approaches, such as decision-analytic modeling, are often necessary in evaluating the comparative clinical utility of medical tests.[8,9]

Herein, we discuss the rationale for and impact of using modeling to assess medical tests for the broad audience of researchers who perform comparative effectiveness reviews and technology assessments, and for policymakers who are debating the merits of different approaches to these products. There is a lot of variation among entities conducting such reviews and assessments in the extent to which they are supportive of, or even familiar with, modeling. We do not explicitly discuss costs and cost-effectiveness analyses, nor do we provide guidelines and recommendations for good modeling practices. Instead, we highlight specific examples to help readers appreciate the role of formal quantitative analyses in the interpretation of evidence on medical tests.

Using Modeling To Interpret Evidence on Medical Test Performance

Putting the Puzzle Together

Studies of test performance give information on the ability of tests to discriminate disease from non-disease. The effect of treatments is usually studied in clinical trials, and the prevalence of disease conditions is typically reported in epidemiological studies. In most instances, one has to integrate evidence from all these types of studies to evaluate the clinical utility of a test in a given setting.[10] For example, the effects of screening for type 2 diabetes on life expectancy have not been directly studied.[11] Instead, there is good evidence on the prevalence of diabetes in specific risk groups, the accuracy of screening, and the downstream effects of proper interventions for diabetes on clinical outcomes.[12] Decision modeling helps explain the implications of screening for impaired glucose tolerance among 45-year-olds with above-average risk and identifies it as a cost-effective approach.[13] Results from ongoing relevant trials are still pending.[12]

Dealing With Uncertainties and Assumptions

Simulation modeling explicitly accounts for uncertainty in key quantities and explicates overt and implicit assumptions.[14] Typically this is done with one- or multi-way sensitivity analyses, where the estimates of one or more model input parameters are systematically varied over prespecified ranges. Alternative modeling options (e.g., comprehensive Bayesian decision analysis[15] and microsimulation models[16]) can incorporate all parameter uncertainties in the model itself.

Tradeoffs

All testing procedures and treatment decisions are associated with benefits, risks, and costs. Decision analysis is a natural framework to assess such tradeoffs. For example, brain biopsy is an invasive procedure that was being considered for the differential diagnosis of suspected herpes simple virus encephalitis in the 1980s. At that time, vidarabine was proven an effective yet toxic treatment for the disease, which has high mortality or long-term neurologic sequelae if it is left untreated. Simultaneously weighing the likelihood of encephalitis and the risks and benefits of brain biopsy and the toxic treatment is extremely challenging, even for seasoned specialists. Decision analyses assessed tradeoffs associated with the choices to give vidarabine empirically, withhold treatment, or biopsy the brain for a diagnosis before initiating treatment, and they provided guidance on thresholds for choosing between the possible options.[17,18]

Comparing Multiple Test-and-Treat Strategies

Often there are many alternative ways to employ existing tests in clinical practice. In such cases, it is not feasible to directly compare all patient management strategies in clinical trials. To attain necessary power, sample sizes become too large, followup duration too long, and costs prohibitively high. Careful modeling offers a feasible alternative. In an evidence report, cost-effectiveness analyses contrasted 17 technologies (tests) and 4 combinations thereof for the diagnosis of acute cardiac ischemia in the emergency department. These had not been compared head-to-head in a clinical trial.[19]

Conversely, management strategies that are deemed promising in modeling analyses could be prioritized for study in actual clinical trials.

Even when assessing a single test, differences in its application can impact on clinical outcomes and associated costs. An example is colonoscopy screening for colorectal cancer. Different start and stop ages for colonoscopy screening, and screening at varying intervals have been proposed and employed. Modeling provides valuable insights on which combinations are optimal. The MISCAN-COLON microsimulation model[16,20] has been used to contrast (among others) screening at different intervals and various start and stop ages.[21] In fact, the U.S. Preventive Services Task Force took into account insights gained from modeling in formulating their screening recommendations.[22] Scrutiny at this level is impossible without simulation modeling.

Succession of Technologies

In fast-paced fields with rapid uptake of novel technologies, continuous innovations can render widely used tests obsolete within a short period of time. By the time of their completion, clinical trials may not be applicable to current standard practice. Examples are the transition from low- to high-resolution computed tomography (CT) and spiral CT, the introduction of magnetic resonance imaging with stronger fields, and the gradual improvement of ultrasound resolution. Careful modeling can help in appreciating the expected benefits, risks, and costs of implementing newer tests by considering improvements in accuracy, as well as potential shifts in the disease spectrum for positive diagnoses.

Exploring Hypothetical Conditions for Diseases With No Effective Treatment

As mentioned before, a test result in itself does not necessarily affect patient-relevant outcomes.[1] This is evident in the case of early diagnosis of a disease for which there is no effective treatment. Notwithstanding patient preferences on how desirable it is to know the result of such a test and the concomitant emotional, cognitive, and behavioral changes conferred by testing and its results, an accurate diagnosis is not expected to impact on patient-relevant outcomes. An attractive way of exploring the clinical utility of such a test is to calculate under what conditions it would be worthwhile to employ it. For example, one can assume that the test would guide the selection of hypothetical treatments with different effectiveness and safety profiles. An evidence report used a decision model to evaluate the ability of positron emission tomography (PET) to guide management of suspected Alzheimer's dementia.[23] Because current medical treatment for Alzheimer's has low efficacy and toxicity, the analysis concluded that routine PET screening is not justified. In fact it was deemed that PET screening becomes attractive only if one assumes that PET would triage patients for treatment with an effective but toxic intervention.

Challenges in Modeling Test-and-Treat Strategies

Models are simplified representations of what can occur in real life, comprehensive enough to capture important behaviors of the simulated scenario and simple enough to study. Problems ensue when models fail to capture important behaviors

(are incomplete or simplistic), because they can mislead. Many excellent publications describe guidelines for good modeling practices, especially in the context of cost-effectiveness analyses.[24-30] Here, we describe methodological and epidemiological considerations that are pertinent to modeling of medical tests. We describe issues that arise when data on important parameters are sparse or unreliable and when test performance is not transferable across studies, and discuss miscellaneous issues that range from statistical considerations to choice of outcomes.

Issues With Insufficient Data

Problems arise when key input quantities of a model are known with low precision or not known at all. Notwithstanding the uniqueness of each case, there are general caveats we can make. From a bird's-eye view and excluding costs from our considerations, simulations of a test-and-treat strategy have at least three groups of important parameters: prevalence of the disease in the setting of interest, test performance and direct effects of testing, and benefits and risks of subsequent treatment(s) in the diseased and nondiseased. Direct effects include testing-induced emotional, cognitive, and behavioral changes or complications of dangerous and invasive tests.

Insufficient or Unreliable Data on Prevalence

Prevalence affects greatly the positive and negative predictive value of a testing strategy. For example, in very low risk populations (very low disease prevalence), even very specific tests can yield relatively large numbers of false positives.[31] When the condition of interest is relatively rare, small absolute changes in prevalence estimates can have great impact in the positive predictive value of a testing strategy.

Valid prevalence estimates are often hard to obtain, especially when one is interested in a particular setting or subpopulation. For example, the prevalence of obstructive sleep apnea among older adults cannot be deduced by the majority of studies of diagnostic tests for sleep apnea, because the latter focus mainly on middle-aged males.[32]

On a related note, many conditions are defined by operational cutoffs along a spectrum of possible clinical presentations. The "disease" is then an arbitrary construct that may or may not correspond to different prognoses. In the sleep apnea example, most published studies defined sleep apnea as $\geq$15 apneas or hypopneas per hour of sleep in a patient with suggestive symptoms and signs.[32] In reality, there is no clinical rationale to distinguish between 13 and 17 apneas or hypopneas per hour of sleep. Yet, when modeling presence or absence of "disease" to examine test-and-treat strategies, such distinctions and simplifications may be unavoidable.[33]

Insufficient or Unreliable Data on Diagnostic Accuracy

A plethora of considerations is relevant here, many of which stem from fundamental shortcomings in the design, conduct, and reporting of diagnostic accuracy studies.[34] The STAndards for the Reporting of Diagnostic accuracy studies (STARD) initiative published a 25-item checklist that aims to improve reporting of studies of diagnostic tests.[35] The reader is referred to the many excellent methodological and empirical explorations that discuss the effects of bias and variation on the performance of medical tests.[36-38]

Here, we opt to discuss in some detail a recurrent challenge that arises when the reference test misclassifies patients in a nontrivial way (tarnished gold standard). Errors made by the reference standard bias the usual estimators of sensitivity and specificity of the index test:[39] They can be underestimates when the results of the two tests are statistically independent, conditional on the true disease status of the patients,[40,41] or overestimates if the results of the two tests are conditionally dependent (i.e., positively correlated either among patients with the disease or among people without the disease[38,42]). For example, in colon cancer diagnosis, both capsule endoscopy (index test) and colonoscopy (reference standard) can be jointly false negative for cancers with little intraluminal manifestation or jointly false positive for some benign intraluminal masses. Treating colonoscopy as an error-free reference standard likely overestimates the ability of the camera pill to detect all colonic cancers.

Insufficient or Poor Data on Effectiveness

In this case, the link between test accuracy and clinical outcomes is weak. Notwithstanding insights gained from modeling of hypothetical treatment effectiveness, as in the aforementioned example on PET and Alzheimer's dementia,[23] it is questionable whether such cases should be routinely subjected to detailed and extensive modeling (at least in the context of interpreting systematic reviews of test performance).[7] Some modeling may still be helpful to identify influential input parameters that must be studied further (e.g., prevalence or effectiveness) and to select the most promising management strategies for diagnostic tests to be further tested in clinical trials.

Transferability or Nontransferability of Diagnostic Performance Across Studies

Studies of medical test performance are not always conducted in the setting of interest[43] and do not necessarily evaluate a test in its anticipated and clinically meaningful role. In simulation models, estimates of sensitivity and specificity are often "borrowed" across settings and roles to make calculations possible. Judgment calls are being made in this process, some of which are discussed below.

Tranferability or Nontransferability of Test Performance Estimates Across Populations and Settings

Estimates of sensitivity and specificity are often considered independently of disease prevalence,[44] and decision analysts typically transfer them across settings with different disease prevalence. However, differences in study inclusion criteria can result in spectrum effects—i.e., differences in the calculated sensitivity and specificity of a medical test as the case-mix of the studied population shifts.[45,46] Indeed, empirical studies have frequently revealed substantial variation of test performance metrics in studies with different disease prevelence.[47]

The transferability of test performance estimates across studies is also influenced by differences in the uptake of a medical test over time or across health systems. Soon after a test gets into practice, health providers may start using it for increasingly broader indications (indication creep), resulting in corresponding shifts in the case-mix of the tested population. Indication creep is not necessarily undesirable as long as there are no changes in the disease spectrum for the positive diagnoses (something that is not easy to

ascertain). However, it should be taken into consideration because it does change the anticipated demand for the technology, and it complicates cost and other projections.

Transferability or Nontransferability of Performance Estimates Across Studies Evaluating the Test in Different Roles

A medical test can have different roles in a test-and-treat strategy, depending on the clinical context. It may be used as the sole diagnostic modality, to triage patients for further workup, or as a confirmatory test for patients selected by prior diagnostic workup. One has to be very cautious in generalizing estimates of diagnostic performance across studies that evaluate a test in different roles. Both the case-mix of tested populations and the positivity thresholds of the test can vary at the same time. For example, a decision analysis compared PET (as the sole diagnostic test) vs. an array of alternative diagnostic strategies for managing patients with solitary pulmonary nodules in their chest radiogram.[48] The decision analysis derived the sensitivity and specificity of PET from studies that used it as a confirmatory test after a positive or inconclusive computed tomography.[48] While this modeling assumption may be defendable, it has to be clearly presented and adequately explored.

Transferability or Nontransferability of Performance Estimates Across Studies Evaluating Different Versions of the Test

Different versions of a test can have very different performance characteristics. Although this will probably be evident to a context expert, it may be missed by modelers who are not intimately familiar with the intricacies of a topic. For example, intact parathyroid hormone (PTH) measurements are used to manage patients with renal osteodystrophy. There are extreme discrepancies between the alternative assays for measuring PTH (from the same manufacturer and other manufacturers)[49] that can result in conflicting recommendations in the same patients. Failure to appreciate such secular trends can render any decision model obsolete and misleading, and can affect real life as well. After all, an unexplained increase in the number of parathyroidectomies in the United States between 1999 and 2002 (coinciding with the transition between assays) has been documented.[50]

Issues With the Choice of Modeling Outcomes

The choice of the outcome that should be maximized—e.g., event-free survival, survival, quality-adjusted life years (QALYs)—depends on the exact key research questions, which also define the perspective of the decision analysis—e.g., patient, health care provider, society. A comprehensive assessment of the value of a medical test should include all patient-relevant benefits and risks related to the duration and quality of the remaining life. Quality-adjusted life expectancy is such a measurement that is easy to understand and that allows comparisons with well-known practices in completely different settings. However, modeling quality-adjusted life expectancy requires information on utilities associated with health states, which are not always available. Alternatively, life expectancy, expected number of health events (e.g., strokes), interventions (e.g., surgeries), or even accuracy in diagnosis and treatment can provide useful information. For example, such outcomes may be used when the time horizon of the simulation does not extend through a lifetime.

Other Issues

Meta-Analysis of Diagnostic Accuracy Data—Which Method?

There are many ways to obtain summary estimates for diagnostic studies,[51-54] and their discussion is outside the scope of this writing. Briefly, separate summaries of sensitivities and specificities ignore the relationship between the two quantities and can result in misleading summaries for both. A simple regression method proposed by Moses and Littenberg[52] calculates a summary receiver operating characteristic (ROC) curve that describes the tradeoff between sensitivity and specificity in diagnostic accuracy studies but is an approximate approach. More rigorous methods are being used increasingly, namely bivariate meta-analyses[53] and hierarchical summary ROC curve analyses.[54] The latter[36-38] methods have been shown to be equivalent in many cases.[55] However, all aforementioned methods rely on a single 2 by 2 table from each study. When modeling explicit thresholds, this is probably excessively wasteful of data, and methods that directly combine ROC curves may be more suitable.

Challenges in the Parameterization and Appraisal of Complex Models

Arbitrarily complex clinical scenarios can be modeled with suitable techniques that include but are not limited to simple trees, Markov models, and microsimulation models. Limitations are posed by data availability or unavailability rather than technical difficulties in implementing simulation approaches.

More advanced modeling can be less transparent and difficult to describe in full technical detail. Increased flexibility often has its toll. Essential quantities may be completely unknown ("deep" parameters) and must be set through assumptions or by calibrating model predictions vs. real empirical data.[56] MISCAN-COLON[16,20] and SimCRC[57] are two microsimulation models describing the natural history of colorectal cancer. Both assume an adenoma-carcinoma sequence for cancer development but differ in their assumptions on adenoma growth rates. Tumor dwell time (an unknown deep parameter in both models) was set to approximately 10 years in MISCAN-COLON[20,58] and to approximately 30 years in SimCRC. Because of such esoteric differences, models can result in different conclusions.

Finally, simulation models should ideally be validated against independent datasets that are comparable to the datasets on which the models were developed.[56] External validation is particularly important for simulation models in which the unobserved deep parameters are set without calibration (based on assumptions and analytical calculations).[16,56]

Final Remarks

By definition, all models are simplified representations of the real world, and therefore incomplete. Exactly for this reason, they are useful. They promote transparency by focusing attention on the influential constituents of each problem, and helping distinguish choices from chances and known parameters from unobserved ones. Modeling facilitates comparisons across testing strategies that have never been, and may never be, contrasted in real life. Formal methodologies for sensitivity analyses help appreciate the impact of uncertainties that accompany parameter estimates. For these

reasons, decision-analytic modeling provides the framework to make informed choices among diagnostic strategies under uncertainty and think through their implications.

The main limitation in performing robust modeling of test-and-treat strategies is the unavailability of good-quality data on key parameters (prevalence of the condition, diagnostic accuracy in the modeled setting, therapeutic efficacy of treatments). All readers of decision analyses should be mindful of the assumptions that are invoked when estimates of sensitivity and specificity are transferred from studies on different settings. Notwithstanding the cautionary notes, we believe that, in the absence of studies comparing test-and-treat strategies with respect to patient-relevant outcomes and provided that good estimates for key parameters can be obtained, decision-analytic modeling should be considered as a standard tool in the assessment of the value of diagnostic tests.

References

1. Bossuyt P, McCaffery K. Patient outcome after testing. Med Decis Making 2009;29(5):E30-E38
2. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844-1847.
3. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144:850-855.
4. Lijmer J, Leeflang M, Bossuyt P. Staged evaluation of medical tests. Med Decis Making 2009;29(5):E13-E21
5. Lord S, Irwig L, Bossuyt P. Evaluating tests: when can comparative evidence of test accuracy and other intermediate outcomes be used as an alternative to randomized controlled trials? Med Decis Making 2009;29(5):E1-E12.
6. Siebert U. When should decision analytic modeling be used in the economic evaluation of health care? Eur J Health Econ 2003;4:143-150.
7. Tatsioni A, Zarin DA, Aronson N, et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005;142:1048-1055.
8. Hunink M, Glasziou P, Siegel J, et al. Decision making in health and medicine. Integrating evidence and values. Cambridge: Cambridge University Press; 2001.
9. Sox H, Blatt M, Higgins M, et al. Medical decision making. American College of Physicians; 2007.
10. Mushlin AI. Challenges and opportunities in economic evaluations of diagnostic tests and procedures. Acad Radiol 1999;6 Suppl 1:S128-S131.
11. Davies MJ, Tringham JR, Troughton J, et al. Prevention of type 2 diabetes mellitus. a review of the evidence and its application in a UK setting. Diabet Med 2004;21:403-414.
12. Waugh N, Scotland G, McNamee P, et al. Screening for type 2 diabetes: literature review and economic modelling. Health Technol Assess 2007;11:iii-xi, 1.
13. Gillies CL, Lambert PC, Abrams KR, et al. Different strategies for screening and prevention of type 2 diabetes in adults: cost effectiveness analysis. BMJ 2008;336(7654):1180-1185.
14. Pauker SG, Kassirer JP. Decision analysis. N Engl J Med 1987;316:250-258.
15. Cooper NJ, Sutton AJ, Abrams KR, et al. Comprehensive decision analytical modelling in economic evaluation: a Bayesian approach. Health Econ 2004;13:203-226.
16. Habbema JD, van Oortmarssen GJ, Lubbe JT, et al. The MISCAN simulation program for the evaluation of screening for disease. Comput Methods Programs Biomed 1985;20:79-93.

17. Barza M, Pauker SG. The decision to biopsy, treat, or wait in suspected herpes encephalitis. Ann Intern Med 1980;92:641-649.

18. Braun P. The clinical management of suspected herpes virus encephalitis. A decision-analytic view. Am J Med 1980;69:895-902.

19. Lau J, Ioannidis JP, Balk EM, et al. Evaluation of technologies for identifying acute cardiac ischemia in emergency departments. Evidence Report/Technology Assessment No. 26. Rockville, MD:Agency for Healthcare Research and Quality; 2001.

20. Loeve F, Boer R, van Oortmarssen GJ, et al. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res 1999;32:13-33.

21. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, et al. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Intern Med 2008;149: 659-69.

22. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2008;149:627-637.

23. Matchar DB, Kulasingam S, Huntington A, et al. Positron emission tomography, single photon emission computed tomography, computed tomography, functional magnetic resonance imaging, and magnetic resonance spectroscopy for the diagnosis and management of Alzheimer's dementia. 2004. Available at: http://www.cms.hhs.gov/coverage/download/id104b.pdf. Accessed November 19, 2009.

24. Decision analytic modelling in the economic evaluation of health technologies. A consensus statement. Consensus Conference on Guidelines on Economic Modelling in Health Technology Assessment. Pharmacoeconomics 2000;17:443-444.

25. Richardson WS, Detsky AS. Users' guides to the medical literature. VII. How to use a clinical decision analysis. B. What are the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA 1995;273:1610-1613.

26. Richardson WS, Detsky AS. Users' guides to the medical literature. VII. How to use a clinical decision analysis. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1995;273:1292-1295.

27. Philips Z, Ginnelly L, Sculpher M, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess 2004;8:iii-xi, 1.

28. Philips Z, Bojke L, Sculpher M, et al. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. Pharmacoeconomics 2006;24:355-371.

29. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models. A suggested framework and example of application. Pharmacoeconomics 2000;17:461-477.

30. Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices--Modeling Studies. Value Health 2003;6:9-17.

31. Meyer KB, Pauker SG. Screening for HIV: can we afford the false positive rate? N Engl J Med 1987;317:238-241.

32. Trikalinos TA, Ip S., Raman G., et al. Home diagnosis of obstructive sleep apnea-hypopnea syndrome. Evidence Report/Technology Assessment. Rockville, MD: Agency for Healthcare Research and Quality; 2007.

33. Trikalinos TA, Lau J. Obstructive sleep apnea-hypopnea syndrome: modeling different diagnostic strategies. Evidence Report/Technology Assessment. Rockville, MD: Agency for Healthcare Research and Quality; 2007.

34. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:W1-W12.

35. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Ann Intern Med 2003;138:40-44.

36. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-1066.

37. Rutjes AW, Reitsma JB, Di NM, et al. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469-476.

38. 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;140:189-202.

39. Hawkins DM, Garrett JA, Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. Stat Med 2001;20:1987-2001.

40. Joseph L, Gyorkos TW. Inferences for likelihood ratios in the absence of a "gold standard". Med Decis Making 1996;16:412-417.

41. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. J Clin Epidemiol 1999;52:943-951.

42. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 2001;57:158-167.

43. Irwig L, Bossuyt P, Glasziou P, et al. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002;324:669-671.

44. Loong TW. Understanding sensitivity and specificity with the right side of the brain. BMJ 2003;327:716-719.

45. Goehring C, Perrier A, Morabia A. Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. Stat Med 2004;23:125-135.

46. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002;137:598-602.

47 Leeflang MMG, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol. 2009;62(1):5-12.

48. Lejeune C, Al ZK, Woronoff-Lemsi MC, et al. Use of a decision analysis model to assess the medicoeconomic implications of FDG PET imaging in diagnosing a solitary pulmonary nodule. Eur J Health Econ 2005;6:203-214.

49. 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;52:1771-1776.

50. Foley RN, Li S, Liu J, et al. The fall and rise of parathyroidectomy in U.S. hemodialysis patients, 1992 to 2002. J Am Soc Nephrol 2005;16:210-218.

51. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics 2003;59:936-946.

52. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;13:313-321.

53. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58:982-990.

54. Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. Acad Radiol 1995;2 Suppl 1:S48-S56.

55. Harbord RM, Deeks JJ, Egger M, et al. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2007;8:239-251.

56. Karnon J, Goyder E, Tappenden P, et al. A review and critique of modelling in prioritising and designing screening programmes. Health Technol Assess 2007;11:iii-xi, 1.

57. National Cancer Institute, Cancer Intervention and Surveillance Modeling Network. Available at: http://cisnet.cancer.gov/. Accessed November 18, 2009.

58. Loeve F, Brown ML, Boer R, et al. Endoscopic colorectal cancer screening: a cost-saving analysis. J Natl Cancer Inst 2000;92:557-563.