Chapter 10

Deciding Whether To Complement a Systematic Review of Medical Tests With Decision Modeling

Thomas A. Trikalinos, M.D., Tufts Evidence-based Practice Center, Boston, MA
Shalini Kulasingam, Ph.D., University of Minnesota School of Public Health, Minneapolis, MN
William F. Lawrence, M.D., M.S., Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, MD

Abstract

Limited by what is reported in the literature, most systematic reviews of medical tests focus on "test accuracy" (or better, test performance) rather than on the impact of testing on patient outcomes. The links between testing, test results, and patient outcomes are typically complex: even when testing has high accuracy, there is no guarantee that physicians will act according to tests results, that patients will follow their orders, or that the intervention will yield a beneficial endpoint. Therefore, test performance is typically not sufficient for assessing the usefulness of medical tests. Modeling (in the form of decision or economic analysis) is a natural framework for linking test performance data to clinical outcomes. We propose that (some) modeling should be considered to facilitate the interpretation of summary test performance measures by connecting testing and patient outcomes. We discuss a simple algorithm for helping systematic reviewers think through this possibility, and illustrate it by means of an example.

Introduction

In this chapter of the *Methods Guide to Medical Test Reviews* (also referred to as the Medical Test Methods Guide) we focus on modeling as an aid to understanding and interpreting the results of systematic reviews of medical tests.[1] Limited by what is reported in the literature, most systematic reviews focus on "test accuracy" (or better, test performance) rather than on the impact of testing on patient outcomes.[2,3] The links between testing, test results, and patient outcomes are typically complex: even when testing has high accuracy, there is no guarantee that physicians will act according to tests results, that patients will follow their orders, or that the intervention will yield a beneficial endpoint.[3] Therefore, test performance is typically not sufficient for assessing the usefulness of medical tests. Instead, one should compare complete test-and-treat strategies (for which test performance is but a surrogate), but such studies are very rare. Most often, evidence on diagnostic performance, effectiveness and safety of interventions and testing, patient adherence, and costs is available from different studies. Much like the pieces of a puzzle, these pieces of evidence should be put together to better interpret and contextualize the results of a systematic review of medical tests.[2,3] Modeling (in the form of decision or

economic analysis) is a natural framework for performing such calculations for test-and-treat strategies. It can link together evidence from different sources; explore the impact of uncertainty; make implicit assumptions clear; evaluate tradeoffs in benefits, harms and costs; compare multiple test-and-treat strategies that have never been compared head-to-head; and explore hypothetical scenarios (e.g., assume hypothetical interventions for incurable diseases).

This chapter focuses on modeling for enhancing the interpretation of systematic reviews of medical test accuracy, and does not deal with the much more general use of modeling as a framework for exploring complex decision problems. Specifically, modeling that informs broader decisionmaking may not fall within the purview of a systematic review. Whether or not to perform modeling for informing decisionmaking is often up to the decisionmakers themselves (e.g., policymakers, clinicians, or guideline developers), who would actually have to be receptive and appreciative of its usefulness.[4] Here we are primarily concerned with a narrower use of modeling, namely to facilitate the interpretation of summary test performance measures by connecting the link between testing and patient outcomes. This decision is within the purview of those planning and performing the systematic review. In all likelihood, it would be impractical to develop elaborate simulation models from scratch merely to enhance the interpretation of a systematic review of medical tests, but simpler models (be they decision trees or even Markov process-based simulations) are feasible even in a short time span and with limited resources.[4–6] Finally, how to evaluate models is discussed in guidelines for good modeling practices,[7–14] but not here.

Undertaking a modeling exercise requires technical expertise, good appreciation of clinical issues, and (sometimes extensive) resources, and should be pursued only when it is likely to be informative. So when is it reasonable to perform decision or cost effectiveness analyses to complement a systematic review of medical tests? We provide practical suggestions in the form of a stepwise algorithm.

A Workable Algorithm

Table 10–1 outlines a practical five-step approach that systematic reviewers could use to decide whether modeling could be used for interpreting and contextualizing the findings of a systematic review of test performance, within time and resource constraints. We outline these steps in an illustrative example at the end of the paper.

Table 10–1: Proposed algorithm to decide if modeling should be a part of the systematic review

Step	Description
1	Define how the test will be used.
2	Use a framework to identify consequences of testing as well as management strategies for each test result.
3	Assess if modeling is useful.
4	Evaluate prior modeling studies.
5	Consider whether modeling is practically feasible in the time frame given.

Step 1. Define how the test will be used.

The PICOTS typology (Population, Intervention, Comparators, Outcomes, Timing, Study design) is a widely adopted formalism for establishing the context of a systematic review.[15] It clarifies the setting of interest (whether the test will be used for screening, diagnosis, treatment guidance, patient monitoring, or prognosis) and the intended role of the medical test (whether it is the only test, an add-on to previously applied tests, or a tool for deciding on further diagnostic

workups). The information conveyed by the PICOTS items is crucial not only for the systematic review, but for planning a meaningful decision analysis as well.

Step 2. Use a framework to identify consequences of testing as well as management strategies for each test result.

Medical tests exert most of their effects in an indirect way. Notwithstanding the emotional, cognitive, and behavioral changes induced by testing and its results,[16] an accurate diagnosis in itself is not expected to affect patient-relevant outcomes. Nor do changes in test performance automatically result in changes in any patient-relevant outcome. From this point of view, test performance (as conveyed by sensitivity, specificity, positive and negative likelihood ratios, or other metrics) is only a surrogate end point. For example, testing for human immunodeficiency virus has both direct and indirect effects. The direct effects could include, but are not limited to, potential emotional distress attributable to the mere process of testing (irrespective of results); the cognitive and emotional benefits of knowing one's carrier status (for accurate results); perhaps the (very rare) unnecessary stress caused by a false positive diagnosis; or possible behavioral changes secondary to testing or its results. Indirect effects include all the downstream effects of treatment choices guided by the test results, such as benefits and harms of treatment in true positive diagnoses, avoidance of harms of treatment in true negative diagnoses, and cognitive and behavioral changes.

Identifying the consequences of testing and its results is a *sine qua non* for contextualizing and interpreting a medical test's (summary) sensitivity, specificity, and other measures of performance. A reasonable start is the analytic framework that was used to perform the systematic review (see the Introduction to this *Medical Test Methods Guide*).[15] This framework can be used to develop a basic tree illustrating test consequences and management options that depend on test results. Going through this exercise helps the reviewers make explicit the clinical scenarios of interest, the alternate (comparator) strategies, and the assumptions made by the reviewers regarding the test-and-treat strategies at hand.

Step 3. Assess whether modeling may be useful.

In most cases of evaluating medical testing, some type of formal modeling will be useful. This is because of the indirectness of the link between testing and health outcomes, and the multitude of test-and-treat strategies that can be reasonably contrasted. Therefore, it may be easier to examine the opposite question (i.e., when formal modeling may not be necessary or useful). We briefly explore two general cases. In the first, one of the test-and-treat strategies is clearly superior to all alternate strategies. In the second, information is too scarce regarding which modeling assumptions are reasonable, what the downstream effects of testing are, or what are plausible values of multiple central (influential) parameters.

The Case Where a Test-and-Treat Strategy Is a "Clear Winner"

A comprehensive discussion of this case is provided by Lord et al.[17,18] For some medical testing evaluations, one can identify a clearly superior test-and-treat strategy without any need for modeling. The most straightforward case is when there is direct comparative evidence for all the test-and-treat strategies of interest. Such evidence could be obtained from well designed, conducted and analyzed randomized trials, or even nonrandomized studies. Insofar as these studies are applicable to the clinical context of interest in the patient population of interest,

evaluate all important test-and-treat strategies, and identify a dominant strategy with respect to both benefits and harms and with adequate power, modeling may be superfluous. In all fairness, direct comparative evidence for all test-and-treat strategies of interest is exceedingly rare.

In the absence of direct comparisons of complete test-and-treat strategies, one can rely on test accuracy only, as long as it is known that the patients who are selected for treatment using different tests will have the same response to downstream treatments. Although the downstream treatments may be the same in all test-and-treat strategies of interest, one *cannot automatically deduce* that patients selected with different tests will exhibit similar treatment response.[3,15,17,18] Estimates of treatment effectiveness on patients selected with one test do not necessarily generalize to patients selected with another test. For example, the effectiveness of treatment for women with early-stage breast cancer is primarily based on cases diagnosed with mammography. Magnetic resonance imaging (MRI) can diagnose additional cases, but it is at best unclear whether these additional cases have the same treatment response.[19] We will return to this point soon.

If it were known that patient groups identified with different tests respond to treatment in the same way, one could select the most preferable test (test-and-treat strategy) based on considerations of test characteristics alone. Essentially, one would evaluate three categories of attributes: the cost and safety of testing; the sensitivity of the tests (ability to correctly identify those with the disease, and thus to proceed to hopefully beneficial interventions); and the specificity of the tests (ability to correctly identify those without disease, and thus avoid the harms and costs of unnecessary treatment). A test-and-treat strategy would be universally dominant if it were preferable versus all alternative strategies and over all three categories of attributes. In case of tradeoffs, i.e., one test has better specificity but another one is safer (with all other attributes being equal), one would have to explore these tradeoffs using modeling.

So how does one infer whether patient groups identified with different tests have (or should have) the same response to treatment? Several situations may be described. Randomized trials may exist suggesting that the treatment effects are similar in patients identified with different tests. For example, the effect of stenting versus angioplasty on reinfarctions in patients with acute myocardial infarction does not appear to differ by the test combinations used to identify the included patients.[20] Thus, when comparing various tests for diagnosing acute coronary events in the emergency department setting, test performance alone is probably a good surrogate for the clinical outcomes of the complete test-and-treat strategies. Alternatively, in the absence of direct empirical information from trials, one could use judgment to infer whether the cases detected from different tests would have a similar response to treatment:

1. Lord et al. propose that when the sensitivity of two tests is very similar, it is often reasonable to expect that the "case mix" of the patients who will be selected for treatment based on test results will be similar, and thus patients would repond to treatment in a similar way.[17,18] For example, Doppler ultrasonography and venography have similar sensitivity and specificity to detect the treatable condition of symptomatic distal deep venous thrombosis.[21] Because Doppler is easier, faster, and non-invasive, it is the preferable test.
2. When the sensitivities of the compared tests are different, it is more likely that the additional cases detected by the more sensitive tests may not have the same treatment response. In most cases this will not be known, and thus modeling would be useful to explore the impact of potential differential treatment response on outcomes. Sometimes we can reasonably extrapolate that treatment effectiveness will be unaltered in the

additional identified cases. This is when the tests operate on the same principle, and the clinical and biological characteristics of the additional identified cases are expected to remain unaltered. An example is computed tomography (CT) colonography for detection of large polyps, with positive cases subjected to colonoscopy as a confirmatory test. Dual positioning (prone and supine) of patients during the CT is more sensitive than supine-only positioning, without differences in specificity.[22] It is very reasonable to expect that the additional cases detected by dual positioning in CT will respond to treatment in the same way as the cases detected by supine-only positioning, especially since colonoscopy is a universal confirmatory test.

The Case of Very Scarce Information

There are times when we lack an understanding of the underlying disease processes to such an extent that we are unable to develop a credible model to estimate outcomes. In such circumstances, modeling is not expected to enhance the interpretation of a systematic review of test accuracy, and thus should not be performed with this goal in mind. This is a distinction between the narrow use of modeling we explore here (to contextualize the findings of a systematic review) and its more general use for decisionmaking purposes. Arguably, in the general decisionmaking case, modeling is especially helpful, because it is a disciplined and theoretically motivated way to explore alternative choices. In addition, it can help identify the major factors that contribute to the uncertainty, as is done in "value of information" analyses.[23,24]

Step 4. Evaluate prior modeling studies.

Before developing a model *de novo* or adapting an existing model, reviewers should consider searching the literature to ensure that the modeling has not already been done. There are several considerations when evaluating previous modeling studies.

First, reviewers need to judge the quality of the models. Several groups have made recommendations on evaluating the quality of modeling studies, especially in the context of cost-effectiveness analyses.[7,9–14] Evaluating the quality of a model is a very challenging task. 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, versus real empirical data.[25] MISCAN-COLON[26,27] and SimCRC[28] are two microsimulation models that describe 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;[27,29] and to approximately 30 years in SimCRC. Because of such differences, models can reach different conclusions.[30] Ideally, simulation models should be validated against independent datasets that are comparable to the datasets on which the models were developed.[25] External validation is particularly important for simulation models in which the unobserved deep parameters are set without calibration, based on assumptions and analytical calculations.[25,26]

Second, once the systematic reviewers deem that good quality models exist, they need to examine whether the models are applicable to the interventions and populations of the current evaluation; i.e., if they match the PICOTS items of the systematic review. In addition, the reviewers need to judge whether methodological and epidemiological challenges have been adequately addressed by the model developers.[3]

Third, the reviewers need to explore the applicability of the underlying parameters of the models. Most importantly, preexisting models will not have had the benefit of the current systematic review to estimate diagnostic accuracy, and they may have used estimates that differ from the ones obtained by the systematic review. Also, consideration should be given to whether our knowledge of the natural history of disease has changed since publication of the modeling study (thus potentially affecting parameters in the underlying disease model).

If other modeling papers meet these three challenges, then synthesizing the existing modeling literature may suffice. Alternatively, developing a new model may be considered, or reviewers could explore the possibility of cooperating with developers of existing high quality models to address the key questions of interest. The U.S. Preventive Services Task Force (USPSTF) and the Technology Assessment program of the Agency for Healthcare Research and Quality (AHRQ) have followed this practice for specific topics. For example, the USPSTF recommendations for colonoscopy screening[31] were informed by simulations based on the aforementioned MISCAN-COLON and SimCRC microsimulation models,[28,32] which were developed outside the EPC program.[26,27]

Step 5. Consider whether modeling is practically feasible in the given time frame.

Even if modeling is determined to be useful, it may still not be feasible to develop an adequately robust model within the context of a systematic review. Time and budgetary constraints, lack of experienced personnel, and other needs may all play a role in limiting the feasibility of developing or adapting a model to answer the relevant questions. Even if a robust and relevant model has been published, it may not necessarily be accessible. Models are often considered intellectual property of their developers or institutions, and they may not be unconditionally available for a variety of reasons. Further, even if a preexisting model is available, it may not be sufficient to address the key questions without extensive modifications by experienced and technically adept researchers. Additional data may be necessary, but they may not be available. Of importance, the literature required for developing or adapting a model does not necessarily overlap with that used for an evidence report.

Further, it may also be the case that the direction of the modeling project changes based on insights gained during the conduct of the systematic review or during the development of the model. Although this challenge can be mitigated by careful planning, it is not entirely avoidable.

If the systematic reviewers determine that a model would be useful but not feasible within the context of the systematic review, consideration should be given to whether these efforts could be done sequentially as related but distinct projects. The systematic review could synthesize available evidence, identify gaps, and estimate many necessary parameters for a model. The systematic review can also call for the development of a model in the future research recommendations section. A subsequent report that uses modeling could provide information on long-term outcomes.

Illustration

Here, we illustrate how the aformentioned algorithm could be applied, using an example of a systematic review of medical tests in which modeling was deemed important to contextualize findings on test performance.[33] Specifically, we discuss how the algorithm could be used to determine if a model is necessary for an evidence report on the ability of positron emission

tomography (PET) to guide the management of suspected Alzheimer's disease (AD), a progressive neurodegenerative disease for which current treatment options are at best modestly effective.[33] The report addressed three key questions, expressed as three clinical scenarios:

1. Scenario A: In patients with dementia, can PET be used to determine the type of dementia that would facilitate early treatment of AD and perhaps other dementia subtypes?
2. Scenario B: For patients with mild cognitive impairment, could PET be used to identify a group of patients with a high probability of AD so that they could start early treatment?
3. Scenario C: Is the available evidence enough to justify the use of PET to identify a group of patients with a family history of AD so that they could start early treatment?

The systematic review of the literature provides summaries of the diagnostic performance of PET to identify AD, but does not include longitudinal studies or randomized trials on the effects of PET testing on disease progression, mortality, or other clinical outcomes. In the absence of direct comparative data for the complete test-and-treat strategies of interest, decision modeling may be needed to link test results to long term patient-relevant outcomes.

Step 1: Define how PET will be used.

The complete PICOTS specification for the PET example is described in the evidence report[33] and is not reviewed here in detail. In brief, the report focuses on the *diagnosis* of the disease (AD) in the three scenarios of patients with suggestive symptoms. AD is typically diagnosed with a clinical examination that includes complete history, physical and neuropsychiatric evaluation, and screening laboratory testing.[34] In all three scenarios, we are only interested in PET as a "confirmatory" test (i.e., we are only interested in PET added to the usual diagnostic workup). Specifically, we assume that PET (1) is used for *diagnosing* patients with different severities or types of AD (mild or moderate AD, mild cognitive impairment, family history of AD), (2) it is an *add-on* to a clinical exam, and (3) it should be compared against the clinical examination (i.e. *no PET* as an add-on test). We are explicitly not evaluating patient management strategies where PET is the only test (i.e., PET "replaces" the typical examination) or where it triages who will receive the clinical examination (an unrealistic scenario). Table 10–2 classifies the results of PET testing.

Table 10–2. Cross-tabulation of PET results and actual clinical status among patients with initial clinical examination suggestive of Alzheimer's

	AD in Long-Term Clinical Evaluation	No AD in Long-Term Clinical Evaluation
PET Suggestive of AD	"True positive"	"False positive"
PET not Suggestive of AD	"False negative"	"True negative"

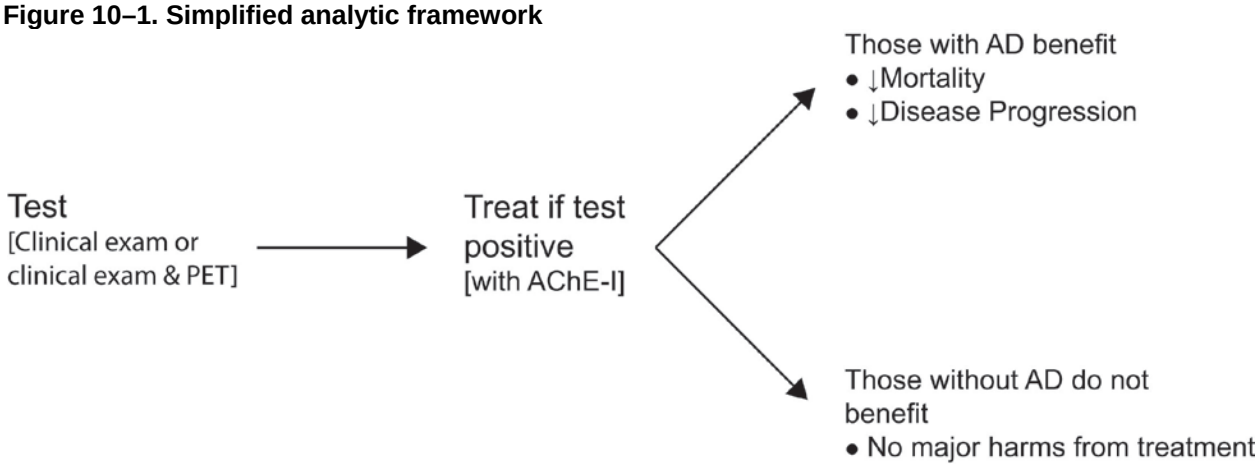
AD = Alzheimer's disease; PET = positron emission tomography
Counts in this table correspond to patients with an initial clinical examination suggestive of AD (as defined in the three clinical scenarios). Patients without suggestive clinical examination are not candidates for PET testing.

Step 2: Create a simplified analytic framework and outline how patient management will be affected by test results.

The PET evidence report does not document any appreciable direct effects or complications of testing with or without PET. Thus, it would be reasonable to consider all direct effects of testing as negligible when interpreting the results of the systematic review of test performance. A

simplified analytic framework is depicted in Figure 10–1, and represents the systematic reviewers' understanding of the setting of the test, and its role in the test-and-treat strategies of interest. The analytic framework also outlines the reviewers' understanding regarding the anticipated effects of PET testing on mortality and disease progression: any effects are only indirect, and conferred exclusively through the downstream clinical decision whether to treat patients. In the clinical scenarios of interest, patients with a positive test result (either by clinical examination or by the clinical examination–PET combination) will receive treatment. However, only those with AD (true positives) would benefit from treatment. Those who are falsely positive would receive no benefit but will still be exposed to the risk of treatment-related adverse effects, and the accompanying polypharmacy. (By design, the evidence report on which this illustration is based did not address costs, and thus we make no mention of costs here.)

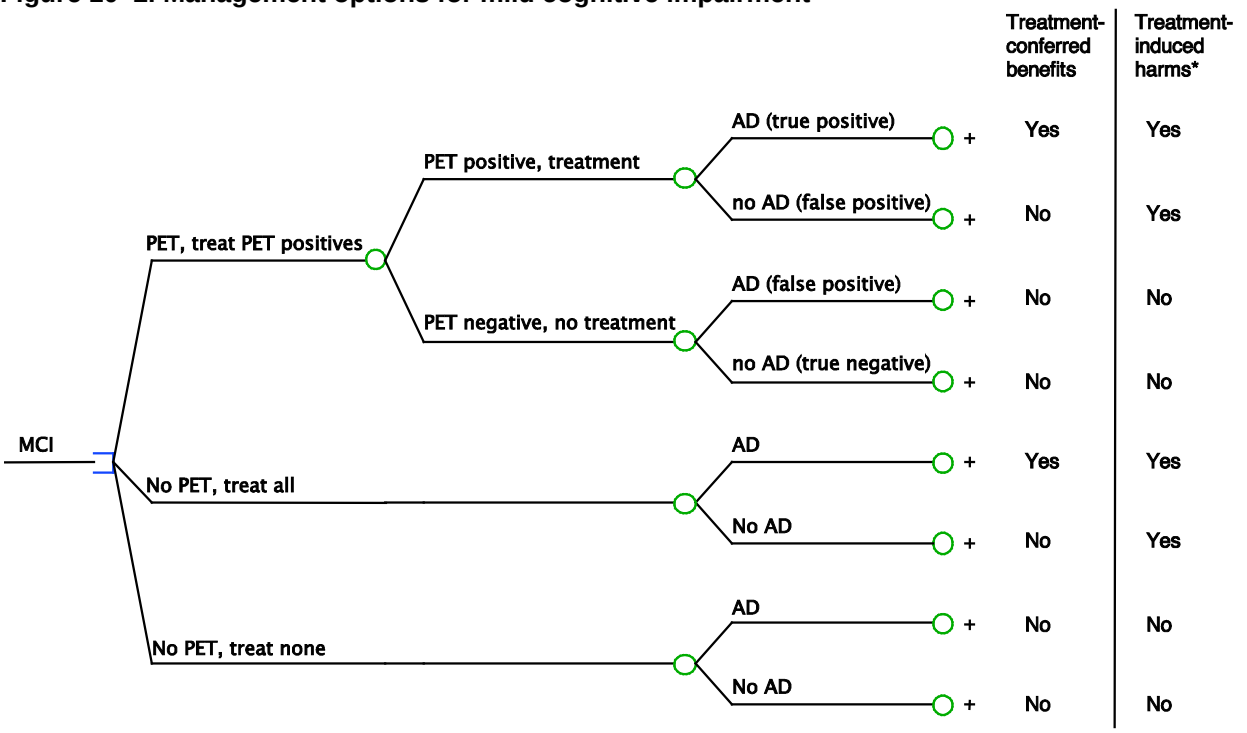
Figure 10–1. Simplified analytic framework



AD = Alzheimer's disease; AChE-I = acetylcholinesterase inhibitors (the treatment available at the time of the evidence report)
The framework assumes no major adverse effects from the treatment.

Figure 10–2 shows an outline of the management options in the form of a simple tree, for the clinical scenario of people with mild cognitive impairment (MCI) in the initial clinical exam (scenario B above). Similar basic trees can be constructed for the other clinical scenarios. The aim of this figure is to outline the management options for positive and negative tests (here they are simple: receive treatment or not) and the important consequences of being classified as a true positive, true negative, false positive or false negative, as well as to make explicit the compared test-and-treat strategies. This simplified outline is an overview of a decision tree for the specific clinical test.

Figure 10–2. Management options for mild cognitive impairment



AD = Alzheimer's disease; MCI = mild cognitive impairment; PET = positron emission tomography
*When applicable. As per the evidence report, the then-available treatment options (acetylcholinesterase inhibitors) do not have important adverse effects. However, in other cases, harms can be induced both by the treatment and the test (e.g., if the test is invasive). The evidence report also modeled hypothetical treatments with various effectiveness and safety profiles to gain insight on how sensitive their conclusions were to treatment characteristics. Note that at the time the evidence report was performed, other testing options for Alzheimer's were not in consideration.

Step 3: Assess whether modeling could be useful in the PET and AD evidence report.

In the example, no test-and-treat strategies have been compared head-to-head in clinical studies. Evidence exists to estimate the benefits and harms of pharmacologic therapy in those with and without AD. Specifically, the treatments for MCI in AD are at best only marginally effective,[33] and it is unknown whether subgroups of patients identified by PET may have differential responses to treatment. Hence, we cannot identify a "clear winner" based on test performance data alone. Thus, modeling was deemed useful here.

Step 4: Assess whether prior modeling studies could be utilized.

In this particular example, the systematic reviewers performed decision modeling. In addition to using the model to better contextualize their findings, they also explored whether their conclusions would differ if the treatment options were more effective than the options currently available. The exploration of such "what if" scenarios can inform the robustness of the conclusions of the systematic review, and can also be a useful aid in communicating conclusions to decisionmakers. It is not stated whether the systematic reviewers searched for prior modeling studies in the actual example. Although we do not know of specialized hedges to identify

modeling studies, we suspect that even simple searches using terms such as “model(s),” “modeling,” “simulat*”, or terms for decision or economic analysis would suffice.

Step 5. Consider whether modeling is practically feasible in the time frame given.

Obviously modeling was deemed feasible in the example at hand.

Overall Suggestions

Many systematic reviews of medical tests focus on test performance rather than the clinical utility of a test. Systematic reviewers should explore whether modeling may be helpful in enhancing the interpretation of test performance data, and in offering insight into the dynamic interplay of various factors on decision-relevant effects.

The five-step algorithm of Table 10–1 can help evaluate whether modeling is appropriate for the interpretation of a systematic review of medical tests.

References

1. Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EC017. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published as a special supplement to the Journal of General Internal Medicine, July 2012.
2. Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005; 142(12 Pt 2):1048-1055.
3. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. Med Decis Making 2009; 29(5):E22-E29.
4. Claxton K, Ginnelly L, Sculpher M, Philips Z, Palmer S. A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme. Health Technol Assess 2004; 8(31):1-103, iii.
5. Meltzer DO, Hoomans T, Chung JW, Basu A. Minimal Modeling Approaches to Value of Information Analysis for Health Research. AHRQ Publication No. 11-EHC062-EF. Rockville, MD: Agency for Healthcare Research and Quality; June 2011. http://ncbi.nlm.nih.gov/books/ NBK62146. Accessed April 10, 2012.
6. Trikalinos TA, Dahabreh IJ, Wong J, Rao M. Future Research Needs for the Comparison of Percutaneous Coronary Interventions with Bypass Graft Surgery in Nonacute Coronary Artery Disease: Identification of Future Research Needs. Future Research Needs Papers No. 1. AHRQ Publication No. 10-EHC068-EF. Rockville, MD: Agency for Healthcare Research and Quality; June 2010. http://ncbi.nlm.nih.gov/ books/NBK51079. Accessed April 10, 2012.
7. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C 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(1):9-17.
8. Trikalinos TA, Balion CM, Colemlan CI, et al. Meta-analysis of test performance when there is a "gold standard." AHRQ Publication No. 12-EHC080-EF. Chapter 8 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

9. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models. A suggested framework and example of application. Pharmacoeconomics 2000; 17(5):461-477.

10. 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(20):1610-1613.

11. 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(16):1292-1295.

12. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess 2004; 8(36):iii-xi, 1.

13. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. Pharmacoeconomics 2006; 24(4):355-371.

14. 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(5):443-444.

15. Matchar DB. Introduction to the Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EHC073-EF. Chapter 1 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

16. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making 2009; 29(5):E30-E38.

17. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009; 29(5):E1-E12.

18. 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(11):850-855.

19. Irwig L, Houssami N, Armstrong B, Glasziou P. Evaluating new screening tests for breast cancer. BMJ 2006; 332(7543):678-679.

20. Nordmann AJ, Bucher H, Hengstler P, Harr T, Young J. Primary stenting versus primary balloon angioplasty for treating acute myocardial infarction. Cochrane Database Syst Rev 2005;(2):CD005313.

21. Gottlieb RH, Widjaja J, Tian L, Rubens DJ, Voci SL. Calf sonography for detecting deep venous thrombosis in symptomatic patients: experience and review of the literature. J Clin Ultrasound 1999; 27(8):415-420.

22. Fletcher JG, Johnson CD, Welch TJ, MacCarty RL, Ahlquist DA, Reed JE et al. Optimization of CT colonography technique: prospective trial in 180 patients. Radiology 2000; 216(3):704-711.

23. Janssen MP, Koffijberg H. Enhancing Value of Information Analyses. Value Health 2009.

24. Oostenbrink JB, Al MJ, Oppe M, Rutten-van Molken MP. Expected value of perfect information: an empirical example of reducing decision uncertainty by conducting additional research. Value Health 2008; 11(7):1070-1080.

25. Karnon J, Goyder E, Tappenden P, McPhie S, Towers I, Brazier J et al. A review and critique of modelling in prioritising and designing screening programmes. Health Technol Assess 2007; 11(52):iii-xi, 1.

26. Habbema JD, van Oortmarssen GJ, Lubbe JT, van der Maas PJ. The MISCAN simulation program for the evaluation of screening for disease. Comput Methods Programs Biomed 1985; 20(1):79-93.

27. Loeve F, Boer R, van Oortmarssen GJ, van BM, Habbema JD. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res 1999; 32(1):13-33.

28. National Cancer Institute, Cancer Intervention and Surveillance Modeling Network. http://cisnet.cancer.gov/colorectal/comparative.html. Accessed April 12, 2012.

29. Loeve F, Brown ML, Boer R, van BM, van Oortmarssen GJ, Habbema JD. Endoscopic colorectal cancer screening: a cost-saving analysis. J Natl Cancer Inst 2000; 92(7):557-563.

30. Zauber AG, Vogelaar I, Wilschut J, Knudsen AB, van Ballegooijen M, Kuntz KM. Decision analysis of coorectal cancer screening tests by age to begin, age to end and screening intervals: Report to the United States Preventive Services Task Force from the Cancer Intervention and Surveillance Modelling Network (CISNET) for July 2007. 2007.

31. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2008; 149(9):627-637.

32. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, van BM, Kuntz KM. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Intern Med 2008; 149(9):659-669.

33. Matchar DB, Kulasingam SL, McCrory DC, Patwardhan MB, Rutschmann OT, Samsa GP et al. Use of positron emission tomography and other neuroimaging techniques in the diagnosis and management of Alzheimer's disease and dementia. AHRQ Technology Assessment, Rockville, MD 2001; http://www.cms.gov/determinationprocess/downloads/id9TA.pdf. Accessed February 6, 2012.

34. Knopman DS, DeKosky ST, Cummings JL, Chui H, Corey-Bloom J, Relkin N et al. Practice parameter: diagnosis of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology 2001; 56(9):1143-1153.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Conflict of interest: None of the authors has any affiliations or financial involvement that conflicts with the information presented in this chapter.

Corresponding author: TA Trikalinos, Tufts Medical Center, 800 Washington St, Box#63, Boston, MA 02111, | Telephone: +1 617 636 0734 | Fax: +1 617 636 8628. Email: Thomas.Trikalinos@tufts.edu.

Suggested citation: Trikalinos TA, Kulasingam S, Lawrence WF. Meta-analysis of test performance when there is a “gold standard.” AHRQ Publication No. 12-EHC082-EF. Chapter 10 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.