Methods Guide for Medical Test reviews

Paper 10

Decision Modeling

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 10. Decision Modeling

Undertaking a modeling exercise requires technical expertise, good appreciation of clinical issues, and (sometimes extensive) resources, and should be pursued when it is maximally informative. In this paper, we provide practical suggestions on how to decide whether modeling is important for interpreting the findings of a systematic review of medical tests using specific examples. We do not discuss *how* to model medical testing and its downstream effects. Many excellent publications describe guidelines for good modeling practices, especially in the context of cost-effectiveness analyses.[1-7]

Common Challenges

Although many medical test evaluations are based on indirect evidence, developing a formal decision model is not always feasible. A commonly encountered challenge is to identify when it is important to perform formal decision or cost-effectiveness analyses to understand the effects of testing on patient-relevant outcomes. The remainder of this paper describes principles for determining when decision modeling is appropriate and feasible.

Principles for Addressing the Challenges

Basic Principle: Adopt a Stepwise Approach

Table 10-1 describes a stepwise approach for assessing whether models are appropriate for reviews of medical tests of different types (e.g., imaging, genetic tests, or strategies combining several tests) in various settings (screening, diagnosis, treatment guidance, prognosis, patient monitoring). Unavoidably, context- and topic-specific considerations will become relevant in practical applications.

Table 10-1. Stepwise approach to determining whether 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 test consequences and management strategy for each test result
3	Assess whether modeling will be useful
4	Evaluate prior modeling studies
5	Consider whether modeling is practically feasible

We expand on each of the above steps in the following sections.

Step 1: Define How the Test Will be Used

As described in the Introduction to this *Medical Test Methods Guide*, the PICOTS typology (Patient population, Intervention, Comparator, Outcomes, Timing, Setting) is a widely adopted

formalism for establishing the context of a systematic review. 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 triages further diagnostic workup). 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 and Management Strategies

Medical tests exert most of their effects in an indirect way. Notwithstanding the emotional, cognitive, and behavioral changes conferred by testing and its results,[8] 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. In principle, test results will 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.

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* Paper 2). This can form the basis for outlining a basic tree illustrating test consequences and management options that depend on test results. This exercise helps reviewers make explicit the clinical scenarios of interest, the alternate (comparator) strategies, and the assumptions they make.

Step 3. Assess Whether Modeling Will 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 reasonably be contrasted. Therefore, it may be easier to consider the opposite question; that is, whether formal modeling will *not* be 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, there is too much uncertainty in multiple central (influential) parameters.

The case of the "clear winner." For some medical testing evaluations, there may be a clearly superior strategy. In the ideal case, the test-and-treat strategies have been directly compared in well-designed and conducted randomized controlled trials (RCTs) or in properly analyzed non-randomized studies. Insofar as these studies match the PICOTS characteristics of interest (i.e., they are applicable to the clinical context of interest in the patient population of interest, evaluate all important test-and-treat strategies, inform outcomes of interest, are conclusive, and are adequately powered), a clear conclusion about the net benefit of alternate strategies could be reached without modeling. This situation is, however, exceedingly rare.

DRAFT

Alternatively, a "clear winner" strategy can be identified only on the basis of test performance when the link between test results and patient-relevant outcomes is very strong. The operating assumption here is that test results dictate the use of an effective treatment or avoidance of a harmful intervention. A complete discussion of this scenario is provided by Lord et al.[9-10] Briefly, a test-and-treat strategy will be dominant when it is better than all other alternate strategies (or at least as good) on the relevant decision criteria (e.g., performance characteristics), and at the same time better than all other strategies in at least some other criteria (e.g., cost or availability). For example, if two tests have identical performance characteristics (sensitivity and specificity, the "decision criteria"), but one of them is cheaper or poses less inconvenience or risk to the patient, the cheaper and safer test should be preferred. For example, Doppler ultrasonography and venography have similar sensitivity and specificity to detect the treatable condition of symptomatic distal deep venous thrombosis;[11] the Doppler test is, however, easier, faster, and non-invasive, and therefore preferable.

In a related scenario, a test may have better performance characteristics than its comparator. Following Lord et al.,[9-10] we distinguish two cases:

1. In the simplest case, the sensitivities are comparable but the specificities differ. In this case, one should pick the test with the better specificity because it avoids the harms of unnecessary further testing or treatment.
2. If the sensitivity of one test is better and the specifities are comparable, then one has to judge whether the additionally detected cases are likely to respond to treatment in the same way as the cases detected by the other test. If this is the case, one can prefer the test with the better sensitivity. If this is not the case, one cannot confidently prefer the test with the better sensitivity, because the additionally detected cases may represent a change in the spectrum of disease, and this in turn affects the effectiveness of downstream interventions.
    a. In the best case, RCTs document that the additionally detected cases respond to treatment. For example, RCTs have shown that estrogen receptor status predicts response to adjuvant tamixifen for breast cancer,[12] suggesting that a strategy that tests for estrogen receptor status to triage treatment is preferable to not testing.
    b. We can reasonably extrapolate that treatment effectiveness will be unaltered in the additional identified cases when the tests operate on the same principle, and the clinical and biological characteristics of the additional identified cases remain unaltered. An (extreme) example is computed tomography colonography for detection of large polyps. Dual positioning (prone and supine) of patients is more sensitive than supine-only positioning, without differences in specificity.[13] It is very reasonable to consider that the additional cases detected by dual positioning will respond to treatment in the same way as the cases detected by supine-only positioning.
3. In all other cases, one has to assess trade-offs in a decision analysis.[9-10]

The case of excessive uncertainty. 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 may not be helpful in illuminating answers to the key questions of the systematic review. Arguably, a modeling exercise that examines many plausible alternatives and conducts wide-ranging sensitivity analyses can still be useful. It can

help explore the major factors that contribute to our uncertainty, should that be considered a goal of the project. In fact, modeling that includes value-of-information analyses may be particularly useful in such cases.[14-15]

Step 4. Evaluate Prior Modeling Studies

Prior to developing a model *de novo*, 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, one has 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.[1-7] 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.[16] MISCAN-COLON[17-18] and SimCRC[19] are two microsimulation models describing the natural history of colorectal cancer. Both assume an adenoma-carcinoma sequence for cancer development, but they 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;[18,20] and to approximately 30 years in SimCRC. Because of such esoteric differences, models can result in different conclusions. Ideally, simulation models should be validated against independent datasets that are comparable to the datasets on which the models were developed.[16] External validation is particularly important for simulation models in which the unobserved deep parameters are set without calibration (based on assumptions and analytical calculations).[16-17]

Second, once the systematic reviewers deem that good-quality models exist, they have to examine whether those models are applicable to the interventions and populations of the current evaluation; that is, whether they match the PICOTS items of the systematic review. In addition, the reviewers have to judge whether methodological and epidemiological challenges have been adequately addressed by the model developers.[21]

Third, reviewers have 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 from the systematic review. Also, consideration should be given to whether 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 existing modeling papers meet these three challenges, then synthesizing this literature may suffice. Alternatively, developing a new model may be considered, or one 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 have followed this practice for specific topics. For example, the USPSTF recommendations for colonoscopy screening[22] were informed by simulations based on the

DRAFT

MISCAN-COLON microsimulation model,[23] which was developed outside the EPC program.[17-18]

Step 5. Consider Whether Modeling is Practically Feasible

Even if a modeling has been determined to be useful, it may still not be feasible to develop a formal model within the context of a systematic review. Time and budgetary constraints, lack of experienced personnel, and other factors may all play a role in limiting the feasibility of developing or adapting a model to answer the relevant questions. 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 readily available or may not be available at all. An important point to note in this context is that the literature required for developing or adapting a model does not necessarily overlap with that used for an evidence report.

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 is not feasible within the context of the systematic review, consideration should be given to whether this effort could be undertaken sequentially as a related but distinct project. The systematic review could summerize available evidence, identify gaps, and and provide estimates of many necessary parameters for a model. The systematic review can also include a call for the development of a model in the future recommendations section. A subsequent report that uses modeling could then inform long-term outcomes.

Illustration

We illustrate the application of the first three steps in determining whether a model should be included in a systematic review with reference to an AHRQ-sponsored report on the ability of positron emission tomograhy (PET) to guide the management of suspected Alzheimer's dementia (AD), a progressive neurodegenerative disease for which current treatment options are at best modestly effective.[24] 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 summarized the diagnotic performance of PET to identify AD, but found no longitudinal studies or RCTs on the effects of PET testing on disease progression, mortality or other clinical outcomes. The reviewers deemed that decision modeling was appropriate to contextualize the information on test performance by linking test results to

long-term patient-relevant outcomes. Modeling was also used to explore whether conclusions would differ if the available treatment options were more effective than the currently available.

In Step 1, reviewers defined how PET would be used. The complete PICOTS specification for the PET example is described in the evidence report[24] and is not reviewed in detail here. In brief, the focus was on *diagnosis* of the disease (AD) among the three scenarios of patients with suggestive symptoms.

AD is typically diagnosed with a clinical exam that includes complete history, physical and neuropsychiatric evaluation, and screening laboratory testing.[25] Reviewers were only interested in PET as a "confirmatory" test; that is, they were interested in adding PET to the usual diagnostic workup. They explicitly did not evaluate patient management strategies where PET is the only test (i.e., PET "replaces" the typical exam) or where PET triages who will receive the clinical exam (an unrealistic scenario). In this particular case, PET is used as an add-on to a clinical exam for *diagnosing* patients with different severities or types of AD (mild or moderate AD, mild cognitive impairment, family history of AD) and is compared against the clinical exam alone (*no PET* as an add-on test). Table 10-2 classifies the results of PET testing.

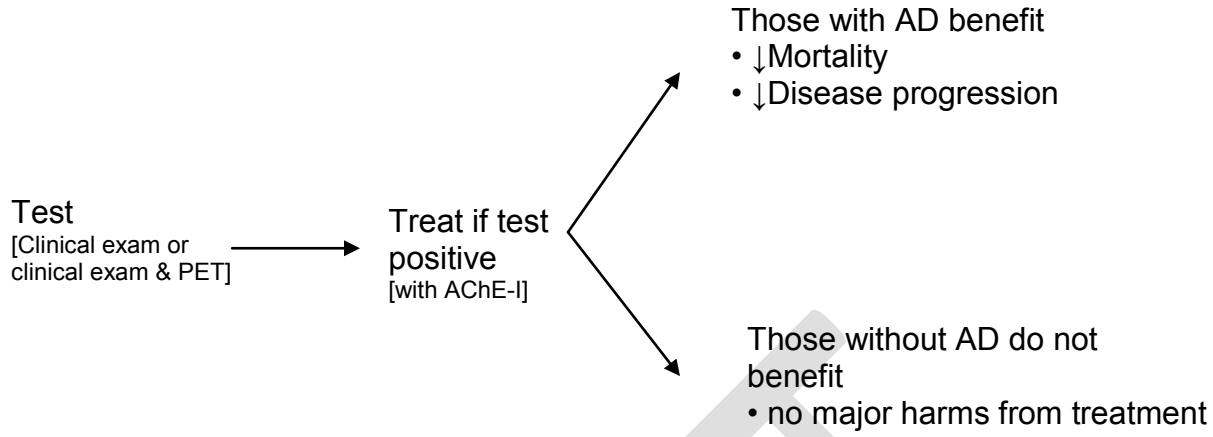
Table 10-2. Cross tabulation of PET results and actual clinical status among patients with initial clinical examination suggestive of Alzheimer's

	AD in long-term clinical evaluation	No AD in long-term clinical evaluation
PET suggestive of AD	"True positive"	"False negative"
PET not suggestive of AD	"False positive"	"True negative"

Abbreviations: AD = Alzheimer's disease; PET = positron emission tomography.
Entries 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.

In Step 2, reviewers further elaborated on the test-and-treat strategies of interest by constructing a simplified analytic framework and outlining patient management options conditional on test results. They assumed that there are no appreciable direct effects or direct complications of testing (with or without PET). A simplified ("straw man") version of the analytic framework is depicted in Figure 10-1. The "straw man" analytic framework conveys that the anticipated effects of PET testing on mortality and disease progression are only indirect and are exclusively conferred through the downstream clinical decision of whether to treat patients. In the clinical scenario of interest, patients with a positive test result (either by clinical exam or by the clinical exam-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.

Figure 10-1. Simplified ("straw man") analytic framework

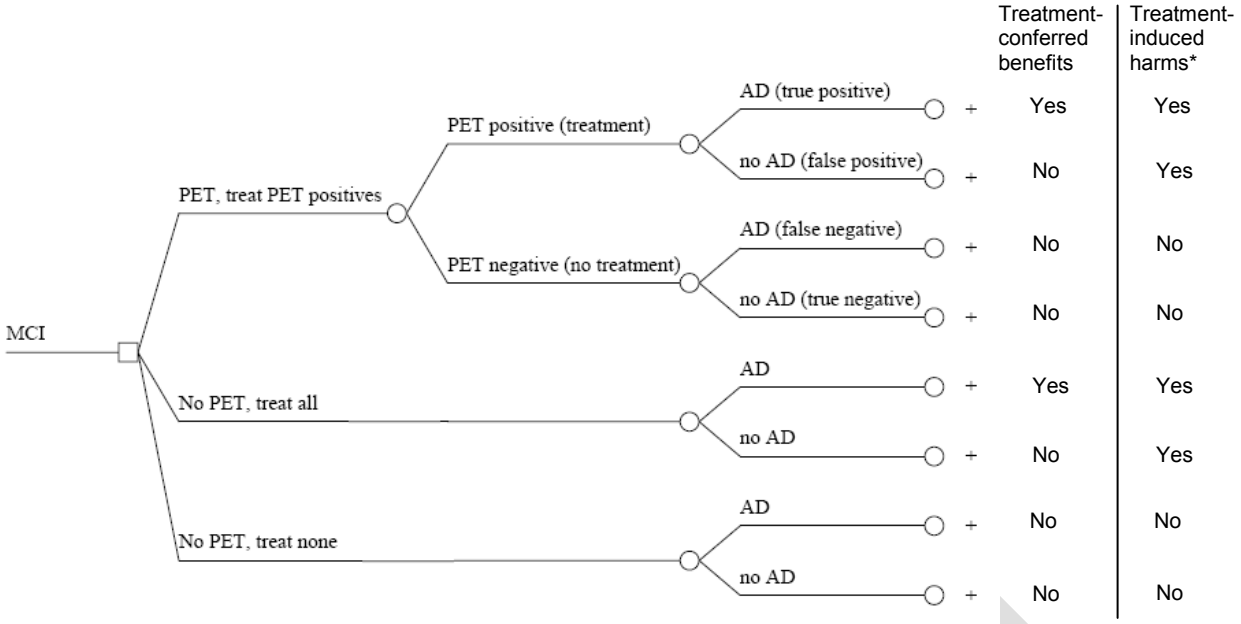


Abbreviations: AD = Alzheimer's disease; AChE-I: acetylcholinestrase inhibitors (the treatment available at the time of the report); PET = positron emission tomography.
The framework assumes no major adverse effects from the treatment.

Figure 10-2 shows an outline of the management options in the form of a simple tree, and for the clinical scenario of people with moderate cognitive impairment (MCI) in the initial clinical exam (scenario B). 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); describe the consequences of being classified as a true positive, true negative, false positive or false negative; and make explicit which test-and-treat strategies are being compared. This simplified outline provides a bird's-eye-view of a decision tree for the specific clinical test.

DRAFT

Figure 10-2. Management options for MCI



*When applicable (harms were not important in the main analyses of the evidence report).
Abbreviations: AD = Alzheimer's disease; MCI = moderate cognitive impairment; PET = positron emission tomography.
As per the evidence report, the then-available treatment options (achetylcholinesterase inhibitors) do not have important adverse effects. However, in other cases, harms can be induced both by the treament 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 are 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 involved assessing whether modeling could be useful. In the PET example, there were no data clearly linking one test-and-treat strategy to improved outcomes over the other strategies, but there are sufficient data to make reasonable estimates of the benefits and harms of pharmacologic therapy in those with and without AD. Conclusions on the superiority of either strategy could not be drawn based only on the performance characteristics of the tests, and therefore a model was developed and analyzed as part of the evidence report. The model offered insights not only on PET but on imaging for AD in general.

Steps 4 and 5 need not be illlustrated.

Summary

Key points are:

- In systematic review of medical tests, it can be illuminating to develop and evaluate a formal decision model that links the decision to use a particular test-and-treat strategy versus appropriate alternate strategies to the likely patient-relevant outcomes.
- In addition to helping identify preferred test strategies, modeling can provide insight into the dynamic interplay of various factors on decision-relevant effects, in turn leading to recommendations for further studies.

- A five-step algorithm can help EPCs evaluate whether modeling is appropriate for the interpretation of a systematic review of medical tests:
  - Step 1. Define how the test will be used.
  - Step 2. Use a framework to identify test consequences and management strategy for each test result.
  - Step 3. Assess whether modeling will be useful.
  - Step 4. Evaluate prior modeling studies.
  - Step 5. Consider whether modeling is practically feasible.

References

1. 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):443-4.

2. Richardson WS, Detsky AS, Evidence-Based Working Group. 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? JAMA 1995;273(20):1610-3.

3. Richardson WS, Detsky AS, Evidence-Based Working Group. Users' guides to the medical literature. VII. How to use a clinical decision analysis. A. Are the results of the study valid? JAMA 1995;273(16):1292-5.

4. 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(36):iii-xi, 1.

5. 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(4):355-71.

6. 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):461-77.

7. 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(1):9-17.

8. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making 2009;29(5):E30-8.

9. 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. Epub2009 Sep 22.

DRAFT

10. Lord SJ, Irwig L, Simes J. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need a randomized trial? Ann Intern Med 2006;144(11):850-5.

11. Gottlieb RH, Widjaja J, Tian L, et al. Calf sonography for detecting deep venous thrombosis in symptomatic patients: experience and review of the literature. J Clin Ultrasound 1999;27(8):415-20.

12. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials Lancet 1998;351(9114):1451-67.

13. Fletcher JG, Johnson CD, Welch TJ, et al. Optimization of CT colonography technique: prospective trial in 180 patients. Radiology 2000;216(3):704-11.

14. Janssen MP, Koffijberg H. Enhancing value of information analyses. Value Health 2009.

15. Oostenbrink JB, Al MJ, Oppe M, et al. Expected value of perfect information: an empirical example of reducing decision uncertainty by conducting additional research. Value Health 2008;11(7):1070-80.

16. 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(52):iii-xi, 1.

17. 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(1):79-93.

18. 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(1):13-33.

19. National Cancer Institute. Cancer Intervention and Surveillance Modeling Network. Available at: http://cisnet.cancer.gov/. Accessed September 4, 2009.

20. Loeve F, Brown ML, Boer R, et al. Endoscopic colorectal cancer screening: a cost-saving analysis. J Natl Cancer Inst 2000;92(7):557-63.

21. 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):E22-9.

22. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2008;149(9):627-37.

23. 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(9):659-69.

DRAFT

24. Matchar DB, Kulasingam SL, McCrory DC, et al. Use of Positron Emission Tomography and other neuroimaging techniques in the diagnosis and management of Altzheimer's disease and dementia. Technology Assessment. Prepared by the Duke Evidence-based Practice Center (under Contract No. 290-97-0014). December 2001. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.cms.hhs.gov/determinationprocess/downloads/id9TA.pdf. Accessed September 4, 2009.

25. Knopman DS, DeKosky ST, Cummings JL, 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):1143-53.

DRAFT