Models in the Development of Clinical Practice Guidelines

J. Dik F. Habbema, PhD; Timothy J. Wilt, MD, MPH; Ruth Etzioni, PhD; Heidi D. Nelson, MD, MPH; Clyde B. Schechter, MD, MA; William F. Lawrence, MD, MS; Joy Melnikow, MD, MPH; Karen M. Kuntz, ScD; Douglas K. Owens, MD, MS; and Eric J. Feuer, PhD

Clinical practice guidelines should be based on the best scientific evidence derived from systematic reviews of primary research. However, these studies often do not provide evidence needed by guideline development groups to evaluate the tradeoffs between benefits and harms. In this article, the authors identify 4 areas where models can bridge the gaps between published evidence and the information needed for guideline development applying new or updated information on disease risk, diagnostic test properties, and treatment efficacy; exploring a more complete array of alternative intervention strategies; assessing benefits and harms over a lifetime horizon; and projecting outcomes for the conditions for which the guideline is intended. The use of modeling as an approach to bridge these gaps (provided that the models are high-quality and adequately validated) is considered. Colorectal and breast cancer screening are used as examples to show the utility of models for these purposes. The authors propose that a modeling study is most useful when strong primary evidence is available to inform the model but critical gaps remain between the evidence and the questions that the guideline group must address. In these cases, model results have a place alongside the findings of systematic reviews to inform health care practice and policy.

Ann Intern Med. 2014;161:812-818. doi:10.7326/M14-0845 www.annals.org
For author affiliations, see end of text.

Clinical practice guidelines facilitate implementation of high-value health care when they are based on consideration of the benefits and harms most relevant to practitioners, their patients, and society. The Institute of Medicine has specified that guidelines should be based on systematic reviews that consider the quality, quantity, and consistency of the relevant evidence (1). This approach is supported internationally (2).

In practice, findings from systematic reviews may not apply directly to the guideline development setting. For example, cancer screening guideline panels may need to determine not only whether they should recommend screening for a specific condition but, if so, the ages at which to start and stop, the frequency, and the test method. For many screening strategies, data to directly address these questions are lacking.

One way to bridge the gap between primary evidence and guideline development is by using models. Models are mathematical frameworks that integrate available data to estimate the health consequences of alternative intervention strategies in patient populations. There are different classes of models with different goals, methodological approaches, or both (3–10). Models have been used to examine the natural history of disease, explain disease occurrence trends, and interrogate harm–benefit tradeoffs of competing policies. Some models express the entire disease process and outcomes at a population level; others (microsimulation models) attempt to construct a virtual population in which persons progress through a disease process.

We focus on modeling to estimate the harm–benefit tradeoffs of different disease management strategies. This requires the existence of a calibrated model of disease progression—that is, a representation of disease progression that is shown to be consistent with observed data. For example, in cancer screening we need a model of disease without screening that yields projections of disease incidence similar to those observed in the absence and presence of screening.

Although the use of models is increasing in guideline development, many guidelines are created without them. In some cases, they are not needed because guideline questions can be adequately addressed by using published primary evidence. However, in many cases, an understanding of how modeling can provide useful information is lacking. This article proposes that models play an important role in integrating and extending the evidence on outcomes of health care interventions. We provide recommendations of when models are likely to be valuable, based on gaps between published research studies and guideline questions. We also discuss aspects of model quality. Finally, we provide direction for how a modeling study should be designed and integrated into the guideline development process.

Examples of Models

We use 2 examples from cancer screening to illustrate our primary points. Screening trials provide primary evidence on benefit but are unable to compare the full range of screening strategies; represent a screening program as it would be broadly implemented; estimate benefits over a lifetime horizon; or fully assess benefits, harms, and costs. To overcome these limitations, models have been used to provide this critical information to support the development of cancer screening recommendations (11).

Colorectal Cancer Screening

Colorectal cancer is one cancer type for which there is broad consensus on screening efficacy. Trials of fecal occult blood tests (FOBTs) have shown significant decreases in colorectal cancer deaths (12). However, disease management and testing technologies have changed since the trials began nearly 40 years ago. For example, new FOBT variants, including Hemoccult SENSA (Beckman Coulter) and immunochemical tests, are available, and use of colonoscopy for screening has increased. No randomized studies of these newer approaches have been conducted,

although estimates of their performance have been published (12).

A study used 2 models to calculate the number of life-years gained (measure of benefits) and the number of diagnostic colonoscopies (measure of harms and resource use) and to compare different screening ages and intervals for available screening tests (13). The models superimposed candidate screening tests with established performance characteristics on representations of adenoma onset, progression to colorectal cancer, and cancer progression. The models reproduced disease incidence trends in published screening trials (13). The results provided evidence for starting screening at age 50 years rather than 40 or 60 years and for stopping at age 75 years rather than 85 years. They supported a 10-year screening interval for colonoscopy and a 1-year interval for high-sensitivity FOBTs. In the models, the Hemoccult II FOBT (Beckman Coulter) had an inferior harm–benefit ratio compared with more recent FOBTs (13). Screening strategies recommended by the U.S. Preventive Services Task Force (USPSTF) were informed by the model results (14).

Mammography Screening

Many mammography screening trials have been conducted worldwide (15). Most indicated that screening reduces breast cancer mortality, but the trials enrolled women of different ages, used varying screening intervals, and had limited numbers of screening rounds. They also used film rather than the newer digital mammography technology and predated contemporary cancer therapies, such as tamoxifen and trastuzumab (15). Therefore, the trials were of limited value for contemporary settings. To inform guideline development (16), a modeling study was used to compare benefits and harms of mammography screening with different starting and stopping ages and screening intervals (17). The models combined previously estimated disease natural history with sensitivity estimates of current mammography tests. Many screening approaches were evaluated. The models indicated that a strategy of biennial mammography for women aged 50 to 69 years maintained an average of 81% of the benefit of annual mammography with half the number of false-positive results. For younger starting ages, the models indicated that initiating biennial screening at age 40 (vs. 50) years reduced mortality by an additional 3%, consumed more resources, and yielded more false-positive results. The USPSTF used the model results to inform its recommendations for biennial screening between ages 50 and 74 years, with individualized decision making before age 50 years and after age 74 years (16).

Which Gaps Between Primary Evidence and Guidelines Can Be Addressed by Using Models?

In both examples, a clear gap existed between the primary evidence from randomized, controlled trials (RCTs) and the evidence needed to develop clinical guidelines, and modeling bridged the gap. The examples help outline 4 areas in which models can be useful (**Table 1**).

Table 1. **Four Areas Where Models Can Bridge the Gap Between Primary Evidence and Guideline Development**

Applying new information on disease risk, prognosis, medical technologies, and treatments to estimate changes in health outcomes
Exploring the full array of alternative intervention strategies
Assessing important benefits and harms over the lifetime of the population
Making appropriate assumptions about attributes of the target population and health care setting for the guideline conditions

Apply New or Updated Information on Disease Risk, Tests, and Treatments

Estimates of mortality benefit from breast cancer screening are derived from trials that started decades ago. However, advances in treatment have decreased disease-specific mortality over this same period, thus potentially reducing benefits of early disease detection. Models can project outcomes by using more recent mortality rates from information about the effect of new treatments on mortality and contemporary life tables (13, 17). In addition, screening tests may change over time. Digital mammography has largely replaced film, and new immunochemical FOBTs have been developed since the initial trials. Models calibrated to disease incidence under older screening technologies can incorporate these newer data sources to provide outcome estimates that are more relevant to current practice.

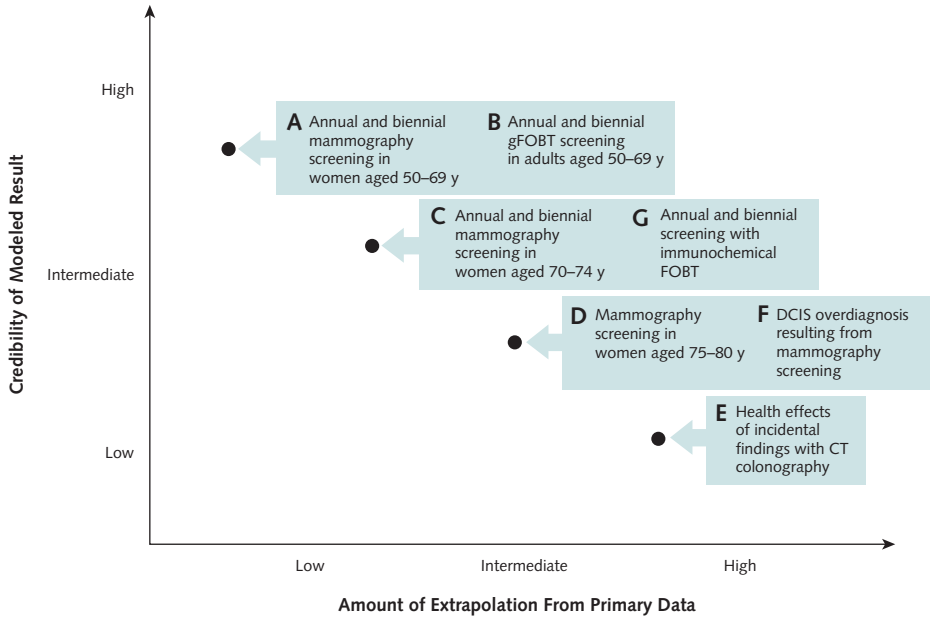
Explore a Wide Range of Possible Intervention Strategies

Many potential screening strategies can be considered that are defined by ages at which to start and stop, intervals, methods, and referral criteria. However, trials can test only a few screening strategies and have limited follow-up. For example, the 7 mammography trials in the systematic review used by the USPSTF had a median of 4 to 5 screening rounds (15), whereas population screening programs often use longer periods. To inform the USPSTF guidelines, the mammography models compared 20 strategies (17). For colorectal cancer, with its many possible screening methods, the models compared 145 strategies (13).

Assess Important Benefits, Harms, and Costs Over the Lifetime of the Population

Models can extend beyond published studies to evaluate new measures of harm–benefit tradeoffs and extrapolate effects on health outcomes beyond the study time horizons. Costs can also be incorporated. Although trials of FOBTs had follow-up exceeding 10 years, they underestimated the lifetime incidence and mortality effects of screening because adenoma detection and removal by colonoscopy after an abnormal FOBT result may prevent colon cancer decades later. Outcomes projected by the colorectal and breast cancer screening models reflected complete screening schedules over lifetime horizons. They included not only the numbers of screening tests, unnec-

Figure. **Illustration of the inverse relationship between the credibility of a modeled result and the required amount of extrapolation from primary data sources.**



Examples A, B, C, D, and G concern mortality reduction and/or life-years gained as the primary modeled outcomes. For examples E and F, the primary modeled outcomes are incidental findings and overdiagnosis, respectively. Examples A and B require relatively little extrapolation from primary data, whereas example E requires extensive extrapolation. The other examples require an intermediate amount of extrapolation. CT = computed tomography; DCIS = ductal carcinoma in situ; FOBT = fecal occult blood test; gFOBT = guaiac fecal occult blood test.

essary biopsies, cancer cases, and cancer deaths prevented but also outcomes difficult to directly measure in trials, including the number of life-years gained and overdiagnosis.

Project Outcomes for the Conditions for Which the Guideline Is Intended

Participants in trials are often highly selected. Specific populations of interest not enrolled or underrepresented in trials may include racial or ethnic minorities, persons younger or older than most trial participants, and those with comorbid conditions. Also, health care conditions and outcomes in settings or countries where trials have been done may differ from those where guidelines apply. For example, prostate cancer incidence rates are higher in the United States than in the United Kingdom, which reflects differences in screening practice (18). However, mortality rates are similar. Models can simulate trial populations and conditions when they are calibrated or validated to these trials (19) and can, in principle, study expected outcomes for populations and settings that differ from those in the trials (13).

How Credible Are Model Results?

Models produce results for many harmful and beneficial outcomes, and it is important for users to know how credible these results are. There is no established scoring system to grade the credibility of a specific model result. Credibility depends on the degree of extrapolation from the primary data needed in calculating the result: More extrapolation leads to less credible results. This is illustrated in the **Figure**, in which various examples express this inverse relationship.

Examples A and B in the **Figure** concern model estimates of favorable effects (life-years gained and cancer deaths prevented) of annual and biennial mammography screening in women aged 50 to 69 years and of guaiac FOBT (gFOBT) screening in adults aged 50 to 69 years. For both mammography and gFOBT screening, several RCTs in the general population (12, 15) inform models of breast and colorectal cancer progression and the effect of early detection on mortality. Most trials had a 2-year screening interval and provided data on interval cancer cases for both a 1-year and a 2-year screening interval. Screening usually focused on persons aged 50 to 69 years for mammography and 45 to 74 years for gFOBT. Therefore, little extrapolation from primary data is required in calculating model results A and B, and they are plotted as having high credibility. Model results are less credible when intervals are longer than those in the trials, screening ages are far outside the range covered by the trial, or outcomes from conditions detected by strategies have not been adequately evaluated.

Example C in the **Figure** concerns annual and biennial mammography in persons aged 70 to 74 years. These ages were included in the enrollment age range in 1 trial and

were included in later screenings in another (16). The example is plotted as having intermediate credibility. Mammography screening in persons aged 75 to 80 years (example D in the **Figure**) is well above the age at enrollment. Larger assumptions are required about the extent to which screening benefits in younger age groups apply to the older age group, so modeling results have lower credibility. The health effects of incidental findings with computed tomography colonography have limited credibility (example E in the **Figure**). No RCTs of screening with computed tomography colonography exist, and deriving credible model estimates for incidental findings would require a much longer follow-up and more accurate assessment of harms and costs than primary data currently provide. Example F in the **Figure** involves overdiagnosis of ductal carcinoma in situ resulting from mammography screening. Overdiagnosis is difficult to estimate from trial data, which explains why results for overdiagnosis of ductal carcinoma in situ vary considerably among models (17). Example G in the **Figure** concerns colorectal cancer screening with immunochemical FOBT. In this case, models need to extrapolate from the RCTs of gFOBT by using the diagnostic test properties of the more recent immunochemical FOBTs, for which no trial exists. As long as reliable information on these test properties is available, it can be used in projections of benefits of the newer tests. The credibility of results for the newer tests will be lower than for gFOBT.

Apart from the amount of extrapolation required, the quality of the primary data to which the models are calibrated is crucial in assessing credibility. Primary data are best when they come from several high-quality trials. This is the case for breast and colorectal cancer screening, the sources of the examples in the **Figure**.

How Can Models Be Used to Weigh Benefits Against Harms and Define Uncertainties?

Models can help guideline development groups quantify the balance of benefits, harms, and costs of intervention strategies and provide an assessment of the uncertainty in the final results. One approach is to consider the outcomes that best capture benefits and harms and to use their ratio to compare candidate policies. This was the USPSTF's approach for colorectal cancer screening, where the number of colonoscopies per life-year gained was used as the primary outcome. This metric captured both the primary benefit (life-years gained) and harms (such as complications associated with colonoscopy and resource use) (14). Similarly, the USPSTF used the number of mammograms per breast cancer death prevented to develop its breast cancer screening recommendations. However, there are no standards for how to determine tradeoffs between benefits and harms or thresholds of acceptable harms per unit of benefit. A well-established tradeoff measure for consideration of costs in economic analysis is the cost-effectiveness ratio (CER), defined as the cost per quality-adjusted life-year (QALY) gained (20). The U.K. National Institute for Health and Care Excellence requests the use of CERs and suggests values of £20 000 or £30 000 per QALY gained as a threshold for evaluation of the net monetary and health benefits of intervention strategies (21). However, there is no clear consensus about a universally acceptable threshold. There are several issues to consider when determining and evaluating CERs and deriving guidelines. First, attaching utility values to all harms and benefits is challenging, may not yield reliable QALY estimates, and may suggest that utilities and preferences vary greatly among patients (22). Second, costs may be difficult to measure and may not be portable across care settings or countries. Third, although some organizations (such as the USPSTF) do not evaluate costs, more than half of the largest U.S. physician societies, including the American College of Physicians, consider the CER and other costs in developing their clinical guidance documents. Among these, approximately half use an explicit mechanism for integrating costs into the strength of recommendations (23). However, CERs are often not available or are not specifically developed for the guideline group. A detailed statement on this matter was recently issued by the American College of Cardiology and American Heart Association (24).

All estimates of harms and benefits are uncertain. The effect of model parameter uncertainty on results can be evaluated by using such methods as sensitivity analysis, probabilistic uncertainty analysis, and scenario analysis (9, 20, 25, 26). In these methods, inputs to and assumptions used in the model are modified to study their effect on outcomes. The modeling studies on breast and colorectal cancer screening performed sensitivity analyses in which test characteristics and other assumptions were varied (13, 17).

What Constitutes a Good Modeling Study?

Once a guideline group has decided that a modeling study would be useful, ensuring that the study is well-conceived and well-conducted is essential (**Table 2**). The model description should include the natural history of the condition and assumptions about how it is modified by the interventions studied. For screening, the way in which the effect of early detection on the primary outcome (which is often age at death for lethal diseases) is modeled is crucial. Model calibration is essential, and the model description should include the natural history parameters and the way they are extracted or estimated. Problems with parameter estimation (especially identifiability, in which different models explain the data equally well) should be recognized and examined in an uncertainty analysis.

The model structure, simulated population, proposed intervention strategies, and projected outcomes should be consistent with the guideline questions. Intervention strategies should include all combinations of test methods and schedules to be modeled, including the management and

Table 2. **Essential Components of a Modeling Study**

Description of the model
Structure
Model inputs and their sources
Time frame
Adaptations for the current application
Intervention strategies that will be modeled
Overview of model assumptions and their justification
Assumptions about the mechanism driving benefit and harm
Assumptions about screening effectiveness
Outcomes to be estimated
Benefits
Harms
Approach to composite outcome measures
Method for balancing harms and benefits
Planned sensitivity analyses to evaluate uncertainty
Study report outline, including the main tables and figures
Study timeline, including interactions with the guideline development group

follow-up of abnormal results. A table of all input values for the natural history parameters and test characteristics, including their sources and the range of uncertainty for each, should be available for review. When input values or their range of uncertainty are not estimable from available data, the way in which they will be set should be stated. A list of underlying model assumptions should accompany the model, including assumptions about access and adherence to the intervention being modeled.

As in randomized trials and observational studies, model quality varies. The value of model applications therefore depends on the quality of the models used. Using no model is better than using a poor one. Although standards for grading the quality of models have not been established, many aspects of model quality exist (**Table 3**) (4, 5, 25, 26). Model quality characteristics include transparency, biological plausibility, verification, calibration, and validation. When important new data become available, the model should be updated and recalibrated. For example, the colorectal cancer models used in the 2008 USPSTF recommendations (14) will need to be recalibrated to fit the results from sigmoidoscopy screening trials (27, 28). How the model has been used in previous applications, its publication record, comparisons with other models, and the modeling team's expertise also provide information about quality. With more experience in the use of models, a grading system for the quality of evidence from modeling studies that mirrors empirical studies may be valuable.

Discussion

The utility of modeling in disease control has long been recognized. Models are used to explore intervention strategies ranging from preventive techniques to salvage therapies across an ever-expanding universe of diseases and conditions. However, modeling is only now beginning to attain legitimacy as part of the guideline development process, and its use in this setting is evolving. This article discusses 4 areas in which guideline groups frequently need to extend results beyond published primary evidence. We show how modeling can be a valuable tool for guideline groups. In addition, we attempt to formalize conditions under which modeling studies are likely to be most useful. The cancer screening application is illustrative; other settings and diseases exist for which modeling can be valuable.

Even when models are used in guideline development, they may not lead to the same recommendations when used by different guideline groups. The USPSTF and the Health Council of the Netherlands both involved modeling teams in the development of their colorectal cancer screening guidelines, and both recommended screening with high-sensitivity FOBT. However, the USPSTF recommended annual screening (14), whereas the Health Council of the Netherlands recommended a biennial interval (29). For cervical cancer screening guidelines, which were informed by models, the USPSTF recommended screening with the Papanicolaou (Pap) test every 3 years from ages 21 to 65 years or the Pap test with human papillomavirus cotesting every 5 years from ages 30 to 65 years (30), whereas the Health Council of the Netherlands recommended Pap screening every 5 years from ages 30 to 60 years or human papillomavirus tests at ages 30, 35, 40, 50, and 60 years (31). Beyond model differences, the types and tradeoffs of harms and benefits considered and the broader policy context may differ. The United States has a history of intensive Pap screening, whereas the Netherlands has a history of less intensive Pap screening (32).

Table 3. **Aspects of Model Quality and Relevance**

Quality
Transparency: Are the structure, parameters, input, and output of the model well-described?
Biological plausibility: Does the natural history implemented in the model make sense in view of current knowledge?
Verification/debugging: Has the model been checked thoroughly to determine that it does what it is intended to do?
Calibration: Has the model been calibrated to high-quality data, and are the statistical methods for parameter estimation and model fitting sound and clearly described?
Validation: What successful prediction tests has the model passed?
Applications record: To what health care decision problems has the model been applied, especially problems similar to the current guideline question?
Scientific record: Has the model resulted in respected, peer-reviewed publications and policy reports?
Collaboration with other models: Has cross-validation with other models taken place, and have joint studies been done?
Track record of the current modeling team members.
Relevance
Is the model able to simulate a population that reflects the target population in terms of age, background incidence of disease, and risk for other-cause death?
Is the model able to project outcomes over the lifetime of the simulated population?
Can the beneficial and harmful outcomes of interest in the relevant population be reliably projected by the model?
Can all competing intervention strategies be implemented in the model?

One approach to make modeling studies more robust is to involve several models. The USPSTF used 2 models for colorectal cancer screening (13), 5 for lung cancer screening (33), and 6 for breast cancer screening (17). Of note, one of the colorectal cancer models gave favorable results for starting screening at age 40 years, contrary to the other model (14). Projections of mortality reduction also varied considerably when the model compared annual versus biennial mammography screening in women aged 50 to 69 years (17). These examples highlight the potential risks of single-model studies, in which results from only 1 of the 2 colorectal cancer models or 1 of the 6 breast cancer models might have been available to the panels. However, multiple-model studies can be more time-consuming and resource-intensive and require explicit steps to ensure that the models produce comparable results. The Cancer Intervention and Surveillance Modeling Network (CISNET) was conceived more than 10 years ago as a prototype for multiple-model studies. On several occasions, guideline development groups have asked CISNET for a joint modeling study (13, 17, 33). The experience and benefits of the CISNET breast cancer model groups working together have been discussed (34), and comparative results have been published (35). When models are available for addressing guideline questions, adapting these models is usually more efficient than developing new ones. Making models public per se without support seldom works because of the expertise required for adapting, testing, and running the model and for postprocessing its output. All studies discussed in this article involved collaborative efforts between the model developers and guideline development groups.

Models can be valuable for specific population subgroups. For colon cancer screening, modeling studies have explored optimal colonoscopy screening policies according to sex and race (36) and family history (37). In the former, it was concluded that the optimal policies did not differ according to sex and race; in the latter, the number of colonoscopies in the optimal strategy varied by a factor of 4 between persons without an affected first-degree relative and those with 2 or more affected first-degree relatives, at least 1 of whom was diagnosed before age 50 years. As personalized medicine becomes more established, the need for this type of modeling will increase.

In conclusion, this article provides a vision of how models can contribute to guideline development when used in addition to systematic reviews of primary evidence. We believe that in many cases models have a place as part of the body of evidence that informs health care practice and policy. Given the tremendous constraints on our ability to conduct large-scale, long-term clinical trials and the need to evaluate unprecedented numbers of new technologies and interventions for disease prevention, detection, and treatment, we anticipate that models will play a growing role in guideline development. Clarifying when it is reasonable to use a model and when a model is likely to be most informative are 2 steps in the evolving science of modeling as part of the policy development process.

From Erasmus MC University Medical Center, Rotterdam, the Netherlands; University of Minnesota School of Medicine, Minneapolis, Minnesota; Fred Hutchinson Cancer Research Center, Seattle, Washington; Oregon Health & Science University, Portland, Oregon; Albert Einstein College of Medicine, New York, New York; Agency for Healthcare Research and Quality, Rockville, Maryland; UC Davis Medical Center, Sacramento, California; Stanford University, Stanford, California; and National Cancer Institute, Bethesda, Maryland.

Disclaimer: This paper reflects the opinions of the authors only and not of guideline development groups in which they are involved; organizations with which they are affiliated; or the Agency for Healthcare Research and Quality, the National Cancer Institute, the U.S. Department of Health and Human Services, or the U.S. Department of Veterans Affairs.

Acknowledgment: The authors thank Therese Miller for her valuable comments.

Financial Support: Dr. Wilt is supported by grants from the U.S. Department of Veterans Affairs and the Agency for Healthcare Research and Quality and contracts to conduct evidence synthesis from the American College of Physicians, the National Kidney Foundation, and Kidney Diseases International. Dr. Nelson is supported by the Agency for Healthcare Research and Quality and the National Institutes of Health. Dr. Schechter is supported by grants from the National Cancer Institute to develop and apply models of breast cancer epidemiology. Dr. Melnikow receives extramural support from the California Health Benefits Review Program and the National Institutes of Health. Dr. Owens is supported by the U.S. Department of Veterans Affairs.

Disclosures: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M14-0845.

Corresponding Author: J. Dik F. Habbema, PhD, Department of Public Health, Erasmus MC University Medical Center, Wytemaweg 80, 3015 CN, Rotterdam, the Netherlands; e-mail, j.d.f.habbema@erasmusmc.nl.

Current author addresses and author contributions are available at www.annals.org.

References

1. **Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, eds.** Clinical Practice Guidelines We Can Trust. Washington, DC: National Academies Pr; 2011.

2. Appraisal of Guidelines for Research and Evaluation (AGREE) II Web site. Accessed at www.agreetrust.org on 24 September 2014.

3. **Garnett GP, Cousens S, Hallett TB, Steketee R, Walker N.** Mathematical models in the evaluation of health programmes. Lancet. 2011;378:515-25. [PMID: 21481448] doi:10.1016/S0140-6736(10)61505-X

4. **Caro JJ, Briggs AH, Siebert U, Kuntz KM; ISPOR-SMDM Modeling Good Research Practices Task Force.** Modeling good research practices—overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Med Decis Making. 2012;32:667-77. [PMID: 22990082]

5. **Roberts M, Russell LB, Paltiel AD, Chambers M, McEwan P, Krahn M; ISPOR-SMDM Modeling Good Research Practices Task Force.** Conceptualizing a model: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-2. Med Decis Making. 2012;32:678-89. [PMID: 22990083]

6. **Siebert U, Alagoz O, Bayoumi AM, Jahn B, Owens DK, Cohen DJ, et al.** State-transition modeling: a report of the ISPOR-SMDM Modeling Good Re-

search Practices Task Force-3. Med Decis Making. 2012;32:690-700. [PMID: 22990084]

7. **Karnon J, Stahl J, Brennan A, Caro JJ, Mar J, Möller J.** Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. Med Decis Making. 2012;32:701-11. [PMID: 22990085]

8. **Pitman R, Fisman D, Zaric GS, Postma M, Kretzschmar M, Edmunds J, et al; ISPOR-SMDM Modeling Good Research Practices Task Force.** Dynamic transmission modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-5. Med Decis Making. 2012;32:712-21. [PMID: 22990086]

9. **Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD; ISPOR-SMDM Modeling Good Research Practices Task Force.** Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. Med Decis Making. 2012;32:722-32. [PMID: 22990087]

10. **Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB; ISPOR-SMDM Modeling Good Research Practices Task Force.** Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Med Decis Making. 2012;32:733-43. [PMID: 22990088]

11. **Stout NK, Dunn MC, Habbema JDF, Levy DT, Feuer EJ.** Using Mathematical Models to Inform Public Policy for Cancer Prevention and Screening. New York: Springer New York; 2013:315-32.

12. **Whitlock EP, Lin JS, Liles E, Beil TL, Fu R.** Screening for colorectal cancer: a targeted, updated systematic review for the U.S. Preventive Services Task Force. Ann Intern Med. 2008;149:638-58. [PMID: 18838718] doi:10.7326/0003-4819-149-9-200811040-00245

13. **Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, van Ballegooijen M, Kuntz KM.** Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Intern Med. 2008;149:659-69. [PMID: 18838717] doi:10.7326/0003-4819-149-9-200811040-00244

14. **U.S. Preventive Services Task Force.** Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2008;149:627-37. [PMID: 18838716] doi:10.7326/0003-4819-149-9-200811040-00243

15. **Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L; U.S. Preventive Services Task Force.** Screening for breast cancer: an update for the U.S. Preventive Services Task Force. Ann Intern Med. 2009;151:727-37. [PMID: 19920273] doi:10.7326/0003-4819-151-10-200911170-00009

16. **U.S. Preventive Services Task Force.** Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2009;151:716-26. [PMID: 19920272] doi:10.7326/0003-4819-151-10-200911170-00008

17. **Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, Draisma G, et al; Breast Cancer Working Group of the Cancer Intervention and Surveillance Modeling Network.** Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. Ann Intern Med. 2009;151:738-47. [PMID: 19920274] doi:10.7326/0003-4819-151-10-200911170-00010

18. **Shibata A, Ma J, Whittemore AS.** Prostate cancer incidence and mortality in the United States and the United Kingdom. J Natl Cancer Inst. 1998;90:1230-1. [PMID: 9719085]

19. **Lansdorp-Vogelaar I, van Ballegooijen M, Boer R, Zauber A, Habbema JD.** A novel hypothesis on the sensitivity of the fecal occult blood test: results of a joint analysis of 3 randomized controlled trials. Cancer. 2009;115:2410-9. [PMID: 19288570] doi:10.1002/cncr.24256

20. **Gold MR, Siegel JE, Russel LB, Weinstein MC, eds.** Cost-effectiveness in Health and Medicine. New York: Oxford Univ Pr; 1996.

21. **National Institute for Health and Care Excellence.** The Guidelines Manual. Appendix G: Methodology checklist: economic evaluations. London: National Institute for Health and Care Excellence; 2009.

22. **Miller AB.** New data on prostate-cancer mortality after PSA screening [Editorial]. N Engl J Med. 2012;366:1047-8. [PMID: 22417259] doi:10.1056/NEJMe1200185

23. **Schwartz JA, Pearson SD.** Cost consideration in the clinical guidance documents of physician specialty societies in the United States. JAMA Intern Med. 2013;173:1091-7. [PMID: 23649494] doi:10.1001/jamainternmed.2013.817

24. **Anderson JL, Heidenreich PA, Barnett PG, Creager MA, Fonarow GC, Gibbons RJ, et al; ACC/AHA Task Force on Performance Measures.** ACC/AHA statement on cost/value methodology in clinical practice guidelines and performance measures: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures and Task Force on Practice Guidelines. Circulation. 2014;129:2329-45. [PMID: 24677315] doi:10.1161/CIR.0000000000000042

25. **Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T, et al.** Modelling in economic evaluation: an unavoidable fact of life [Editorial]. Health Econ. 1997;6:217-27. [PMID: 9226140]

26. **Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, et al; ISPOR Task Force on Good Research Practices—Modeling Studies.** 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. [PMID: 12535234]

27. **Atkin WS, Edwards R, Kralj-Hans I, Wooldrage K, Hart AR, Northover JM, et al; UK Flexible Sigmoidoscopy Trial Investigators.** Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. Lancet. 2010;375:1624-33. [PMID: 20430429] doi:10.1016/S0140-6736(10)60551-X

28. **Segnan N, Armaroli P, Bonelli L, Risio M, Sciallero S, Zappa M, et al; SCORE Working Group.** Once-only sigmoidoscopy in colorectal cancer screening: follow-up findings of the Italian Randomized Controlled Trial—SCORE. J Natl Cancer Inst. 2011;103:1310-22. [PMID: 21852264] doi:10.1093/jnci/djr284

29. **Health Council of the Netherlands.** A national colorectal cancer screening programme. Publication no. 2009/13E. The Hague, the Netherlands: Health Council of the Netherlands; 2009.

30. **Moyer VA; U.S. Preventive Services Task Force.** Screening for cervical cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2012;156:880-91. [PMID: 22711081] doi:10.7326/0003-4819-156-12-201206190-00424

31. **Health Council of the Netherlands.** Population screening for cervical cancer. Publication no. 2011/07E. The Hague, the Netherlands: Health Council of the Netherlands; 2011.

32. **Habbema D, De Kok IM, Brown ML.** Cervical cancer screening in the United States and the Netherlands: a tale of two countries. Milbank Q. 2012;90:5-37. [PMID: 22428690] doi:10.1111/j.1468-0009.2011.00652.x

33. **de Koning HJ, Meza R, Plevritis SK, ten Haaf K, Munshi VN, Jeon J, et al.** Benefits and Harms of Computed Tomography Lung Cancer Screening Programs for High-Risk Populations. AHRQ publication no. 13-05196-EF-2. Rockville, MD: Agency for Healthcare Research and Quality; 2013.

34. **Habbema JD, Schechter CB, Cronin KA, Clarke LD, Feuer EJ.** Modeling cancer natural history, epidemiology, and control: reflections on the CISNET breast group experience. J Natl Cancer Inst Monogr. 2006:122-6. [PMID: 17032902]

35. **Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, et al; Cancer Intervention and Surveillance Modeling Network (CISNET) Collaborators.** Effect of screening and adjuvant therapy on mortality from breast cancer. N Engl J Med. 2005;353:1784-92. [PMID: 16251534]

36. **Lansdorp-Vogelaar I, van Ballegooijen M, Zauber AG, Boer R, Wilschut J, Winawer SJ, et al.** Individualizing colonoscopy screening by sex and race. Gastrointest Endosc. 2009;70:96-108, 108.e1-24. [PMID: 19467539] doi:10.1016/j.gie.2008.08.040

37. **Wilschut JA, Steyerberg EW, van Leerdam ME, Lansdorp-Vogelaar I, Habbema JD, van Ballegooijen M.** How much colonoscopy screening should be recommended to individuals with various degrees of family history of colorectal cancer? Cancer. 2011;117:4166-74. [PMID: 21387272] doi:10.1002/cncr.26009

Current Author Addresses: Dr. Habbema: Department of Public Health, Erasmus MC University Medical Center, Wytemaweg 80, 3015 CN, Rotterdam, the Netherlands.
Dr. Wilt: Minneapolis VA Center for Chronic Disease Outcomes Research and the University of Minnesota School of Medicine, 1 Veterans Drive (111-0), Minneapolis, MN 55417.
Dr. Etzioni: Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B230, PO Box 19024, Seattle, WA 98109-1024.
Dr. Nelson: Pacific Northwest Evidence-based Practice Center, Oregon Health & Science University, Mailcode BICC, 3181 SW Sam Jackson Park Road, Portland, OR 97239.
Dr. Schechter: Department of Family & Social Medicine, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461.
Dr. Lawrence: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850.
Dr. Melnikow: Center for Healthcare Policy and Research, University of California, Davis, 4860 Y Street, Sacramento, CA 95817.
Dr. Kuntz: School of Public Health, University of Minnesota, 420 Delaware Street SE, D360, Mayo MMC 729, Minneapolis, MN 55455.
Dr. Owens: Center for Health Policy/Center for Primary Care and Outcomes Research, Stanford University, 117 Encina Commons, Stanford, CA 94305.
Dr. Feuer: Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive, Bethesda, MD 20892.

Author Contributions: Conception and design: J.D.F. Habbema, R. Etzioni, H.D. Nelson, W.F. Lawrence, J. Melnikow, K.M. Kuntz, D.K. Owens, E.J. Feuer.
Analysis and interpretation of the data: J.D.F. Habbema, T.J. Wilt, R. Etzioni, C.B. Schechter, J. Melnikow, K.M. Kuntz, D.K. Owens, E.J. Feuer.
Drafting of the article: J.D.F. Habbema, T.J. Wilt, R. Etzioni, H.D. Nelson, C.B. Schechter, J. Melnikow, D.K. Owens, E.J. Feuer.
Critical revision of the article for important intellectual content: J.D.F. Habbema, T.J. Wilt, R. Etzioni, H.D. Nelson, C.B. Schechter, W.F. Lawrence, J. Melnikow, K.M. Kuntz, D.K. Owens, E.J. Feuer.
Final approval of the article: J.D.F. Habbema, R. Etzioni, H.D. Nelson, C.B. Schechter, W.F. Lawrence, J. Melnikow, K.M. Kuntz, D.K. Owens, E.J. Feuer.
Provision of study materials or patients: C.B. Schechter.
Statistical expertise: R. Etzioni, C.B. Schechter, K.M. Kuntz.
Administrative, technical, or logistic support: E.J. Feuer.
Collection and assembly of data: J.D.F. Habbema, R. Etzioni.