Volume 13 • Number 5 • 2010
VALUE IN HEALTH

Empirically Evaluating Decision-Analytic Models

Jeremy D. Goldhaber-Fiebert, PhD,[1,2] Natasha K. Stout, PhD,[2,3] Sue J. Goldie, MD, MPH[2]

[1]Stanford Health Policy, Centers for Health Policy and Primary Care and Outcomes Research, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA; [2]Program in Health Decision Science, Department of Health Policy and Management, Harvard School of Public Health, Boston, MA, USA; [3]Department of Population Medicine, Harvard Medical School/Harvard Pilgrim Health Care, Boston, MA, USA

ABSTRACT

Objectives: Model-based cost-effectiveness analyses support decision-making. To augment model credibility, evaluation via comparison to independent, empirical studies is recommended.

Methods: We developed a structured reporting format for model evaluation and conducted a structured literature review to characterize current model evaluation recommendations and practices. As an illustration, we applied the reporting format to evaluate a microsimulation of human papillomavirus and cervical cancer. The model's outputs and uncertainty ranges were compared with multiple outcomes from a study of long-term progression from high-grade precancer (cervical intraepithelial neoplasia [CIN]) to cancer. Outcomes included 5 to 30-year cumulative cancer risk among women with and without appropriate CIN treatment. Consistency was measured by model ranges overlapping study confidence intervals.

Results: The structured reporting format included: matching baseline characteristics and follow-up, reporting model and study uncertainty, and stating metrics of consistency for model and study results. Structured searches yielded 2963 articles with 67 meeting inclusion criteria and found variation in how current model evaluations are reported. Evaluation of the cervical cancer microsimulation, reported using the proposed format, showed a modeled cumulative risk of invasive cancer for inadequately treated women of 39.6% (30.9–49.7) at 30 years, compared with the study: 37.5% (28.4–48.3). For appropriately treated women, modeled risks were 1.0% (0.7–1.3) at 30 years, study: 1.5% (0.4–3.3).

Conclusions: To support external and projective validity, cost-effectiveness models should be iteratively evaluated as new studies become available, with reporting standardized to facilitate assessment. Such evaluations are particularly relevant for models used to conduct comparative effectiveness analyses.

Keywords: cancer, cost-effectiveness, methods, simulation model, validation.

Introduction

Decision-makers must choose among policies in the face of considerable uncertainty about their future benefits and costs. Although empirical studies provide information on specific interventions, they often report intermediate outcomes and assess a limited number of strategies. Delaying a decision while waiting for optimal data on long-term outcomes may not be feasible. Decision-analytic computer models—applied in cost-effectiveness analyses—can provide valuable insights to guide interim decision-making. By synthesizing available information and formally considering uncertainty, models extrapolate short-term study results to project long-term outcomes of different policy choices [1–3].

For policymakers to incorporate modeling into their decisions, models must be transparent and credible. It is therefore important to appreciate both the benefits and limitations of decision models. Though they incorporate the best available data and science, models are inherently simplified representations of the real world. As such, inaccuracies in their predictions can result when the modeled outcomes of a policy depend on features that have been simplified. To promote transparency and credibility, standardization of methods (e.g., using a "reference case") and model quality assessment have been advocated, especially because models are incorporated into comparative effectiveness and guidelines development processes [3,4]. To further enhance transparency and credibility, comparison of model-projected outputs to outcomes from independent, empirical studies—termed "model evaluation"—is an important undertaking. Further, because policy recommendations incorporating model-derived guidance are used long after the initial model-based analyses, continued, iterative evaluation of models is warranted.

We developed a structured reporting format for model evaluation and undertook a literature review of model evaluation as currently used in medicine and public health. The review showed that although model evaluation is recommended in guidelines and review articles, more detailed guidance regarding how best to undertake such an evaluation is limited. We also reviewed published applications of model evaluation as currently practiced, finding a wide range of methods and reporting formats. Such heterogeneity limits comparisons and hinders efforts to include results from multiple studies in comparative effectiveness analyses and policy formulation.

To illustrate the use of the reporting format, we applied it to our ongoing evaluation of an empirically calibrated human papillomavirus (HPV) and cervical cancer microsimulation model used for policy analyses in the United States [5–8]. With policy debates surrounding prevention of HPV infection and cervical cancer evolving rapidly, iterative model evaluation is particularly appropriate.

Methods

Definitions

Our study focuses on external model evaluation, which refers here to external consistency or projective validity. Model evaluation is defined as the comparison—without adjustment of the model's parameters—of the model's outcomes to outcomes from empirical study data not used in the model's construction, parameterization, or calibration [3,9]. In contrast, calibration is defined as the process of determining the values of unobservable

Address correspondence to: Jeremy D. Goldhaber-Fiebert, Stanford Health Policy, Centers for Health Policy and Primary Care and Outcomes Research, 117 Encina Commons, Stanford, CA 94305-6019, USA. E-mail: jeremygf@stanford.edu

10.1111/j.1524-4733.2010.00698.x

Table 1 Structured model evaluation report

Reporting category	Detailed information
Empirical study description	Rationale for selection (e.g., large, high-quality, only available)
	Design (e.g., randomized controlled trial, observational cohort)
	Relevant details (e.g., sample size, year conducted, geographical location)
	Explicit statement that the study was not used to construct the model
Baseline characteristics	Characteristics used to match the modeled and actual study populations (e.g., ages and, potentially, birth cohorts, risk factors)
	Statement if study does not provide sufficient information such as the distribution of baseline characteristics or, more likely, the co-occurrence of risk factors
	Statement if the model does not incorporate certain risk factors thought to be influential in the studied outcome
	Statement that the model does not explicitly match on baseline characteristics as both the model and study are generally representative of similar populations
Study protocol	How subjects are identified, enrolled, and, potentially, assigned to exposure or treatment
	Follow-up and variability of follow-up
	Loss to follow-up
	How study measurements are performed (tests used)
	Statement if study does not provide sufficient data on variability and loss to follow-up and any assumptions used
Study outcomes	Point estimates
	Measures of uncertainty
Model outcomes	Point estimates
	Measures of uncertainty
	How model uncertainty was generated (e.g., probabilistic sensitivity analysis, empirical calibration)
Model consistency	What metrics of consistency were used
	How model and study outcomes meet these metrics
	Assessment of robustness and/or possible reasons for differences
	Likely impact on policy conclusions

parameters by constraining model output to replicate observed data [10]—explicitly modifying model parameters to match such data.

Recommended Reporting of External Model Evaluation

Recommendations are summarized in a structured reporting format for model evaluation in Table 1. When there are multiple studies that a model's output could be compared with, a statement describing the decision rule by which studies were selected should be included along with relevant information about the studies' designs and populations. A statement characterizing the match between the modeled population, and baseline study characteristics and protocol should be included. Model outputs may differ from external studies both because of the model's parameters and structure, and because of differences in the baseline characteristics of the population modeled or the modeled follow-up protocol. By matching baseline characteristics and follow-up, evaluation can focus on key differences (i.e., parameters and structure) relevant for policy analyses. Outcomes and associated uncertainty bounds should be reported for both the study and the model. The use of multiple outcomes from multiple studies, if available, can add further credibility. Model and study outcomes should be compared using a metric of consistency that is explicitly stated. Assessment of similarities and differences between modeled and study outcomes should be assessed along with their potential impact on long-term policy-relevant outcomes. In situations where model evaluation is reported as part of a policy analysis or when the model is compared with large numbers of studies [11,12], space constraints may preclude reporting all information in the article itself. A supplemental Web appendix to describe fully the model's comparison with each study should be included.

Structured Literature Review

We characterized existing guidance for and practice of model evaluation in the context of model-based cost-effectiveness analyses, performing a structured literature review. The goals of the review were: 1) to synthesize any prior guidance on model evaluation and report dissenting opinions in the literature; and 2) to document whether guidance, in the absence of reporting standards, is being followed by simulation modelers. Our searches of MEDLINE (1966–2008) used keyword combinations including: "model," "validation," "evaluation," "external," "simulation," "policy," "cost," "effectiveness," "utility," "decision," "analytic," "health," and "economic." Detailed descriptions of the specific searches performed are included in Appendix SA and can be found at: http://www.ispor.org/Publications/value/ViHsupplementary/ViH13i5_Goldhaber-Fiebert.asp. The bibliographies of resulting articles were also scanned for additional sources.

We limited inclusion of articles returned by our search procedure to those that are in English, about human health, and simulated disease processes with discrete event simulation, Markov, semi-Markov, microsimulation, or differential equation models. Additionally, articles were also required to address external model evaluation—the comparison of the model to empirical studies not used in the model's construction or calibration [3,9].

The search procedure is presented in Figure 1. Searches yielded 2963 articles. Title/abstract reviews resulted in 238 selected for full text review. Among these 238 articles, 67 articles (23 reviews/guidelines and 44 applications) met all inclusion criteria. The remaining articles were excluded because they only cited prior evaluation (49 articles); only mentioned the need for external evaluation (30 articles); or for other less common reasons such as only assessing internal consistency (i.e., comparing model outputs to data used in its construction), or modeling animal diseases.

Results

Current Guidance

In the 23 review and guidelines articles referencing external model evaluation, a small minority framed model evaluation as a validation exercise concerned with the "falsification" of the model [13,14]. The majority—especially more recent reviews and

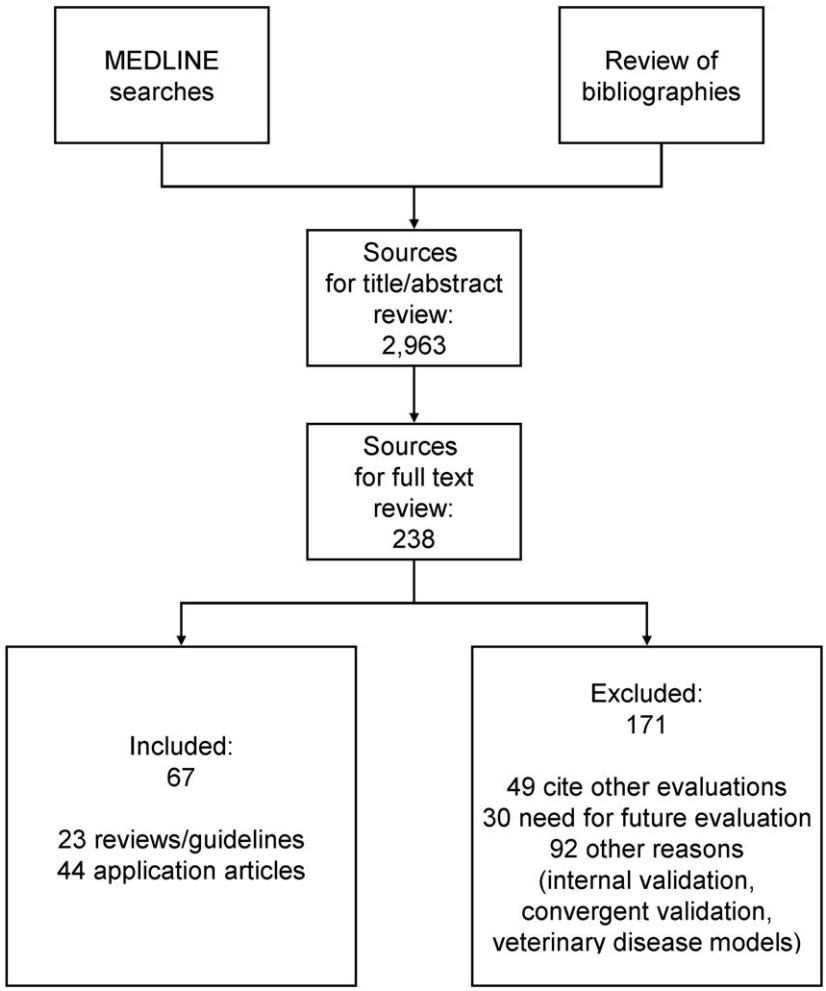

Figure 1 The search procedures used to identify review and guidelines articles as well as application articles dealing with external model evaluation in the context of disease simulation models of human health.

guidelines—view model evaluation in terms of a cumulative process that adds to the general credibility of results through the comparison of the model to independent data not used in its construction [4,15–34].

Although the rationale for comparing models and data, and the types of comparisons made are comprehensive, relatively few articles [15,16,29] provide best practices for performing model evaluations. The following general practices, included in our structured reporting format, were consistent with the 23 reviews: 1) simulation of the same distribution of baseline characteristics and study protocol when generating model outcomes; 2) comparison to multiple studies; 3) comparison to multiple outcome types; and 4) incorporation of model and study uncertainty in the comparison of outcomes. Notably, few articles mentioned criteria for assessing the consistency of model-predicted outcomes and study results. Although some suggested the use of statistical comparisons, others focused on assessing consistency based on the impact on policy conclusions [29,33], presumably by updating the model and its parameters, and reassessing policies recommended for the decision problem.

Current Practice

We assessed 44 application articles identified via the literature review, evaluating whether the general recommendations contained in guidelines and review articles were in common practice [6,11,12,35–75]. Of these 44, 25% compared results only to randomized controlled trials, 66% compared results only to observation studies and population registries, and 9% compared results to both types of studies.

We found that 84% (37 of 44) of articles reported some baseline characteristic (e.g., age distribution, birth year, starting risk level, etc.) and that 66% (29 of 44) of articles explicitly reported matching the study protocol. For comparisons to population-representative observational studies (e.g., statistics from comprehensive cancer registries), baseline characteristics and surveillance may have been assumed to be implicitly similar and, hence, not reported. Consequently, we considered reporting of baseline characteristics and protocol for those articles making comparisons to randomized controlled trials. In this subgroup, we found that 73% reported both matching baseline characteristics and protocol, whereas 53% of comparisons to observational studies reported both baseline and protocol.

We found that 55% (24 of 44) of articles compared model outcomes to outcomes from multiple studies, though this likely underestimates the true percentage because evaluations of the same model could have been published in several, separate articles. Furthermore, even when modelers may have wished to compare with multiple studies, lack of data may have impeded them. Models and studies were compared based on multiple outcomes in 93% (41 of 44) of articles.

Application articles used a wide variety of assessment metrics in making their comparisons of study and model outcomes. Comparisons included: 1) the relative or absolute difference in model and study point estimates; 2) the overlap of model outcomes with study uncertainty ranges; and 3) formal statistical tests where a *P*-value greater than 0.05 or other critical level implied no detectable statistical difference. Less than half of the articles, 43% (19 of 44), reported on uncertainty when comparing model and study outcomes. Reporting uncertainty took a variety of forms, including the use of confidence intervals, or ranges of study and modeled outcomes, or the use of statistical tests. Only 11% (5 of 44) of the articles provided estimates of uncertainty about modeled outcomes, with most comparing the point estimates of the model to the study results.

Example Application: US Cervical Cancer Model

Context and model. Cervical cancer is caused by infection with high-risk types of HPV [76]. Systematic, high-quality screening programs can prevent cervical cancer [77–80]. With the availability of alternative screening technologies [81] and two prophylactic HPV vaccines targeting common oncogenic HPV subtypes [82], important policy questions are raised.

Previously, we constructed an individual-based simulation model of the natural history of HPV and cervical cancer. The model was empirically calibrated to multiple studies. Empirical calibration indentified multiple sets of model inputs with which to generate uncertainty ranges around modeled policy outcomes [6]. Briefly, the microsimulation tracks individual females from age 9. Each month, there is an age-specific risk of infection with HPV, with HPV stratified by risk type. Infections can be cleared, potentially leading to type-specific immunity. Infections can also persist, potentially causing cervical intraepithelial neoplasia (CIN). CIN is classified as low grade (CIN1) or medium to high grade (CIN2,3). For women who do not clear their infections, productive high-risk infection accompanying persistent CIN2,3 can progress to invasive cervical cancer. All women face a monthly age-specific risk of death from other causes, and women with invasive cancer face an additional cancer stage-specific mortality risk.

Prior use and model evaluation. The model has been used in policy analyses [7,8]. As the model was built and calibrated in the absence of interventions (i.e., a natural history model), we focused our iterative external model evaluation on the addition of screening and treatment to prevent cervical cancer, specifically examining CIN2,3 (high-grade CIN), which is particularly important as it is often the primary threshold for treatment and has highly uncertain transition rates to invasive cancer. Previously, we compared the model to studies of the overall impact of screening and treatment of high-grade precancer on population outcomes, such as cancer incidence reduction, finding reasonable consistency [6]. Subsequently, we compared the model to studies of the relationship between cancer and prior frequency of screening and to those that directly tested the ability of screening that included HPV deoxyribonucleic acid (DNA) testing to detect prevalent and incident high-grade CIN [5], again finding reasonable consistency.

Current external model evaluation. The details of the current external model evaluation are presented in Table 2, following our structured reporting format. The current evaluation involved a comparison of the model to a unique, retrospective cohort study of the cumulative risk of cervical cancer for women with high-grade CIN who were either adequately or inadequately treated over 30 years of follow-up [83]. To simulate the study, baseline data on women's age group and birth cohort as well as cohort-specific New Zealand life tables [84] were used. Cumulative risk of cervical cancer at 5, 10, 20, and 30 years post-enrollment were noted for the two groups (adequate and inadequate treatment). Model uncertainty depended on parameter uncertainty identified with the model's original empirical calibration [6,7]. Consistency was assessed based on overlap of the study's 95% confidence intervals and model ranges. The modeled outcomes were largely consistent with the study at all time points with substantial overlap between study confidence intervals and model ranges. Further details on the current step of the external model evaluation are available in Appendix SB at: http://www.ispor.org/Publications/value/ViHsupplementary/ViH13i5_Goldhaber-Fiebert.asp.

Conclusions

Model evaluation—comparing study and model-based outcomes—is an important way to enhance transparency and credibility for models used in policy analyses. In order for models to inform comparative effectiveness analyses and guidelines setting efforts, credibility must be established for the models used. Model evaluation plays an important role in this endeavor. Our structured literature review of 67 articles revealed strong support for model evaluation and some commonalities in approach. We also found a degree of heterogeneity in how results of model evaluations are reported. To promote transparency and credibility in model evaluation, we developed a structured reporting format and demonstrated its utility by applying it to an iterative assessment of a previously published microsimulation of HPV and cervical cancer.

Our findings illustrate the value of a concise reporting format such as the one we suggest in iterative model evaluations. Given the limited space available in peer-reviewed publications, identifying and reporting key details of model evaluations facilitate transparency for readers. The use of such a reporting format in conjunction with longer Web appendixes enables both standardized reporting and full, critical assessment. In our example, the suggested reporting format conveys both overall consistency between the HPV and cervical cancer model and the empirical study, and summarizes the manner in which the model evaluation was conducted. An important job for analysts is to continue to assess their models in comparison to new studies as they are published. Our model's credibility is built gradually through iterative comparison to and consistency with empirical studies as they become available. Finally, though studies may comment on different populations from those represented by the original model, analysts can still exploit both similarities and differences to learn about their models, determining parameter changes needed to equalize baseline population characteristics and then comparing model and study outcomes using similar follow-up protocols. In our comparison of a US model to a study from New Zealand, we were able to make reasonable comparisons, adjusting for background mortality and the age distribution of high-grade precancer.

Our suggested reporting format and structured literature review have several limitations. First, our literature review uses multiple keyword searches in MEDLINE combined with scans of article bibliographies and does not include a full systematic review of other databases such as EMBASE, the use of Medical Subject Heading terms, or searches outside of the medicine and public health domains. Our search results do, however, identify application articles covering a broad range of journals, years,

Table 2 Cervical cancer application of the structured reporting format

Reporting category	Detailed information
Empirical study description	The study provides external model evaluation of the model's progression component for high-grade CIN (CIN3) to invasive cervical cancer. The study used is a unique, retrospective cohort study of women in New Zealand of adequately and inadequately treated CIN3 over a 30-year period published by McCredie and colleagues [83]. Model evaluation focuses on two groups of women from the study: 1) 92 women with treatment deemed inadequate who had only a punch or wedge biopsy; and 2) 299 women with treatment deemed adequate. The study was not used in the construction, parameterization, or calibration of the model.
Baseline characteristics	Model and study were matched on baseline age group and birth cohort data in which cohort-specific New Zealand life tables [84] were used. All women in the study and model had high-grade CIN at baseline. Other potentially influential risk factors including subtype of HPV infection and history of prior infections were unavailable for matching.
Study protocol	The study involved initial treatment of high-grade CIN for some of the participants. All participants were then followed for up to 30 years with cumulative incidence of invasive cancer reported at 5, 10, 20, and 30 years. In the model-simulated study, all women are followed for 30 years as the study did not detail loss to follow-up.
Study outcomes	Inadequate treatment for high-grade CIN, cumulative cancer incidence: 5 years: 17.4% (95% CI 11.1–26.9) 10 years: 26.2% (95% CI 18.4–36.5) 20 years: 34.0% (95% CI 25.2–44.7) 30 years: 37.5% (95% CI 28.4–48.3) Adequate treatment for high-grade CIN, cumulative cancer incidence: 5 years: 0.0% 10 years: 0.3% (95% CI 0.1–2.4) 20 years: 1.5% (95% CI 0.4–3.3) 30 years: 1.5% (95% CI 0.4–3.3)
Model outcomes	Inadequate treatment for high-grade CIN, cumulative cancer incidence: 5 years: 15.5% (95% CI 12.2–19.4) 10 years: 27.8% (95% CI 22.2–34.5) 20 years: 37.5% (95% CI 29.7–46.8) 30 years: 39.6% (95% CI 30.9–49.7) Adequate treatment for high-grade CIN, cumulative cancer incidence: 5 years: 0.0% 10 years: 0.1% (95% CI 0.1–0.2) 20 years: 0.6% (95% CI 0.4–0.7) 30 years: 1.0% (95% CI 0.7–1.3) The study protocol was simulated multiple times using the parameter uncertainty derived from the model's previous empirical calibration to generate the range of uncertainty around the simulated cumulative risk estimates.
Model consistency	Consistency was assessed based on overlap of the study's 95% CI and model ranges. Point estimates for adequately treated women differed by 0.5% at 30 years with at least 90% of model parameter sets consistent with study CIs at all time points. Point estimates for inadequately treated women differed by no more than 3.4% across the 30-year period with at least 80% of model parameter sets consistent with study CIs. Assessment of robustness involved the effects of all-cause death rates used; regression rates from high-grade CIN to no CIN; and the influence of prior infections and immunity. Results remained consistent with the model for death rates, and prior infections and immunity, suggesting that current high-grade CIN strongly determined future cancer risk and survival. Regression rates from high-grade CIN were highly influential. Although our policy model used rates of high-grade CIN progression and regression calibrated to prevalence data, the current model evaluation supports the credibility of the modeled reduction in cancer risk and subsequent survival gain from adequate treatment for screen-detected and diagnosed high-grade CIN.

CI, confidence interval; CIN, cervical intraepithelial neoplasia; HPV, human papillomavirus.

diseases, and research groups. Second, although our literature review finds heterogeneity in how model evaluations are reported, we cannot comment on the reasons for these differences. For example, although our review finds that less than half of application articles report uncertainty, we cannot distinguish simple lack of availability from decisions not to include such information. Our suggested reporting format elicits this type of information explicitly.

Our study highlights a general methodological issue—how to appropriately describe consistency. Although some view model evaluation in terms of formal "validation" through attempts at "falsification," the general consensus is that repeated comparisons in which model and study outcomes are consistent with one another add to the credibility of model-projected long-term health and economic outcomes. In light of this, we recommend a description of consistency that highlights both model and study uncertainty, and comments on their similarity in reference to policy-relevant outcomes. One example of such an approach would be to report the study and model point estimates, and the degree of overlap between study confidence intervals and model uncertainty ranges.

An important question raised in the context of model evaluation is what to do when modeled outcomes are inconsistent with study outcomes. As the model represents the synthesis of available data and expertise, inconsistency represents an opportunity to understand why there are differences and improve the model. One potential reason for inconsistency is a lack of sufficient detail reported in the real-world studies used in the model evaluation. For example, without individual-level data from the empirical study, the co-occurrence of baseline characteristics and their relationship to patterns of follow-up are difficult to determine. It is unclear whether and how much bias is introduced when simulating such studies without this detail, an area meriting further study. Therefore, inconsistency with respect to a study should not immediately result in discarding the model or its findings. Nevertheless, caution should be used in applying the model in the domain that the study comments upon. For example, if the study elaborates the effect of a particular drug on a particular pathway or intermediate outcome, then modeled long-term outcomes based on this drug, pathway, or intermediate outcomes should be further critically examined. When possible, comparisons should be made to other studies with results relevant in that area. The

primary goal should be to understand the mechanism and magnitude of divergence and its likely impact on policy conclusions, based on model-projected outcomes.

Model evaluation is an important and feasible part of model-based cost-effectiveness analyses for many diseases and policy areas. We anticipate expansion in the use of model-based analyses as interest in comparative effectiveness grows. As model-based analyses increasingly incorporate methodological best practices, there is optimism that decision-makers will further utilize model-projected results. Improving the manner in which models and their analytic approaches are disseminated will aid in this regard. To augment the transparency and credibility of long-term, model-projected costs and benefits, iterative assessment of a model's consistency with external data from published studies should be undertaken when possible and reported with sufficient information for others to assess.

Source of financial support: This work was supported in part by the National Cancer Institute (R01 CA093435). Jeremy Goldhaber-Fiebert was the recipient of the National Science Foundation's Graduate Research Fellowship. Natasha Stout was supported by the National Cancer Institute (F32 CA125982).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix SA. Structured Review, queries and resulting items.

Appendix SB. External Model Evaluation of the HPV/Cervical Cancer Model, detailed description of procedures and results for the iterative of assessment of the HPV/Cervical Cancer microstimulation model in comparison to a long-term longitudinal study.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

1 National Institute of Health and Clinical Excellence. Our Guidance. 2009. Available from: http://www.nice.org.uk/ [Accessed August 1, 2009].
2 Edejer TT-T, World Health Organization. Making Choices in Health: WHO Guide to Cost-Effectiveness Analysis. Geneva: World Health Organization, 2003.
3 Gold MR. Cost-Effectiveness in Health and Medicine. New York: Oxford University Press, 1996.
4 Chilcott J, Brennan A, Booth A, et al. The role of modelling in prioritising and planning clinical trials. Health Technol Assess 2003;7:1–125.
5 Giorgi Rossi P, Zappa M. Re: cost-effectiveness of cervical cancer screening with human papillomavirus DNA testing and HPV-16,18 vaccination. J Natl Cancer Inst 2008;100:1654, author reply 54–5.
6 Goldhaber-Fiebert JD, Stout NK, Ortendahl J, et al. Modeling human papillomavirus and cervical cancer in the United States for analyses of screening and vaccination. Popul Health Metr 2007; 5:11.
7 Goldhaber-Fiebert JD, Stout NK, Salomon JA, et al. Cost-effectiveness of cervical cancer screening with human papillomavirus DNA testing and HPV-16,18 vaccination. J Natl Cancer Inst 2008;100:308–20.
8 Stout NK, Goldhaber-Fiebert JD, Ortendahl JD, et al. Trade-offs in cervical cancer prevention: balancing benefits and risks. Arch Intern Med 2008;168:1881–9.
9 Hunink MGM. Decision Making in Health and Medicine: Integrating Evidence and Values. Cambridge and New York: Cambridge University Press, 2001.
10 Stout NK, Knudsen AB, Kong CY, et al. Calibration methods used in cancer simulation models and suggested reporting guidelines. Pharmacoeconomics 2009;27:533–45.
11 Eddy DM, Schlessinger L. Validation of the archimedes diabetes model. Diabetes Care 2003;26:3102–10.
12 Palmer AJ, Roze S, Valentine WJ, et al. Validation of the CORE Diabetes Model against epidemiological and clinical studies. Curr Med Res Opin 2004;20(Suppl. 1):S27–40.
13 Ramsey SD, McIntosh M, Etzioni R, et al. Simulation modeling of outcomes and cost effectiveness. Hematol Oncol Clin North Am 2000;14:925–38.
14 Sornette D, Davis AB, Ide K, et al. Algorithm for model validation: theory and applications. Proc Natl Acad Sci U S A 2007;104:6562–7.
15 American Diabetes Association Consensus Panel. Guidelines for computer modeling of diabetes and its complications. Diabetes Care 2004;27:2262–5.
16 The Mount Hood 4 Modeling Group. Computer modeling of diabetes and its complications: a report on the Fourth Mount Hood Challenge Meeting. Diabetes Care 2007;30:1638–46.
17 Bagust A, McEwan P. Guidelines for computer modeling of diabetes and its complications: response to American Diabetes Association Consensus Panel. Diabetes Care 2005;28:500, author reply 500–1.
18 Drummond MF. Health economic models: a question of balance—summary of an open discussion on the pharmacoeconomic evaluation of non-steroidal anti-inflammatory drugs. Rheumatology (Oxford) 2000;39(Suppl. 2):29–32.
19 Hammerschmidt T, Goertz A, Wagenpfeil S, et al. Validation of health economic models: the example of EVITA. Value Health 2003;6:551–9.
20 Hay JW. Evaluation and review of pharmacoeconomic models. Expert Opin Pharmacother 2004;5:1867–80.
21 Institute of Medicine (US) Committee for Evaluating Medical Technologies in Clinical Use, Institute of Medicine (US) Division of Health Sciences Policy, Institute of Medicine (US) Division of Health Promotion and Disease Prevention. Assessing Medical Technologies. Washington, DC: National Academy Press, 1985.
22 Kim SY, Goldie SJ. Cost-effectiveness analyses of vaccination programmes: a focused review of modelling approaches. Pharmacoeconomics 2008;26:191–215.
23 Kohn MC. Achieving credibility in risk assessment models. Toxicol Lett 1995;79:107–14.
24 Marshall DA, Douglas PR, Drummond MF, et al. Guidelines for conducting pharmaceutical budget impact analyses for submission to public drug plans in Canada. Pharmacoeconomics 2008;26:477–95.
25 McCabe C, Dixon S. Testing the validity of cost-effectiveness models. Pharmacoeconomics 2000;17:501–13.
26 Newall AT, Beutels P, Wood JG, et al. Cost-effectiveness analyses of human papillomavirus vaccination. Lancet Infect Dis 2007;7:289–96.
27 Oreskes N. Evaluation (not validation) of quantitative models. Environ Health Perspect 1998;106(Suppl. 6):1453–60.
28 Russell LB. Modelling for cost-effectiveness analysis. Stat Med 1999;18:3235–44.
29 Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models. A suggested framework and example of application. Pharmacoeconomics 2000;17:461–77.
30 Taylor-Robinson D, Milton B, Lloyd-Williams F, et al. Policymakers' attitudes to decision support models for coronary heart disease: a qualitative study. J Health Serv Res Policy 2008;13: 209–14.
31 Unal B, Capewell S, Critchley JA. Coronary heart disease policy models: a systematic review. BMC Public Health 2006;6:213.
32 Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices—Modeling Studies. Value Health 2003;6:9–17.

33 Weinstein MC, Toy EL, Sandberg EA, et al. Modeling for health care and other policy decisions: uses, roles, and validity. Value Health 2001;4:348–61.
34 Hodges JS, Dewar JA, Rand Corporation, et al. Is It You or Your Model Talking?: A Framework for Model Validation. Santa Monica, CA: Rand, 1992.
35 Arveux P, Wait S, Schaffer P. Building a model to determine the cost-effectiveness of breast cancer screening in France. Eur J Cancer Care (Engl) 2003;12:143–53.
36 Baker RD. Use of a mathematical model to evaluate breast cancer screening policy. Health Care Manag Sci 1998;1:103–13.
37 Borg S, Glenngard AH, Osterborg A, et al. The cost-effectiveness of treatment with erythropoietin compared to red blood cell transfusions for patients with chemotherapy induced anaemia: a Markov model. Acta Oncol 2008;47:1009–17.
38 Brown T, Peerapatanapokin W. The Asian Epidemic Model: a process model for exploring HIV policy and programme alternatives in Asia. Sex Transm Infect 2004;80(Suppl. 1):i19–24.
39 Chang CM, Lin WC, Kuo HS, et al. Estimation and prediction system for multi-state disease process: application to analysis of organized screening regime. J Eval Clin Pract 2007;13:867–81.
40 Chen TH, Kuo HS, Yen MF, et al. Estimation of sojourn time in chronic disease screening without data on interval cases. Biometrics 2000;56:167–72.
41 Cook JR, Yin D, Alemao E, et al. Development and validation of a model to project the long-term benefit and cost of alternative lipid-lowering strategies in patients with hypercholesterolaemia. Pharmacoeconomics 2004;22(Suppl. 3):37–48.
42 Custer B, Johnson ES, Sullivan SD, et al. Community blood supply model: development of a new model to assess the safety, sufficiency, and cost of the blood supply. Med Decis Making 2005;25:571–82.
43 Dasbach EJ, Insinga RP, Yang YC, et al. The cost-effectiveness of a quadrivalent human papillomavirus vaccine in Taiwan. Asian Pac J Cancer Prev 2008;9:459–66.
44 Dedes KJ, Szucs TD, Imesch P, et al. Cost-effectiveness of trastuzumab in the adjuvant treatment of early breast cancer: a model-based analysis of the HERA and FinHer trial. Ann Oncol 2007;18:1493–9.
45 Etzioni R, Gulati R, Falcon S, et al. Impact of PSA screening on the incidence of advanced stage prostate cancer in the United States: a surveillance modeling approach. Med Decis Making 2008;28:323–31.
46 Grover SA, Coupal L, Paquet S, et al. Cost-effectiveness of 3-hydroxy-3-methylglutaryl-coenzyme A reductase inhibitors in the secondary prevention of cardiovascular disease: forecasting the incremental benefits of preventing coronary and cerebrovascular events. Arch Intern Med 1999;159:593–600.
47 Grover SA, Coupal L, Zowall H, et al. Cost-effectiveness of treating hyperlipidemia in the presence of diabetes: who should be treated? Circulation 2000;102:722–7.
48 Grover SA, Paquet S, Levinton C, et al. Estimating the benefits of modifying risk factors of cardiovascular disease: a comparison of primary vs secondary prevention. Arch Intern Med 1998;158: 655–62.
49 Hoffmann T, Brunner H. Model for simulation of HIV/AIDS and cost-effectiveness of preventing non-tuberculous mycobacterial (MAC)-disease. Eur J Health Econ 2004;5:129–35.
50 Holman RR, Retnakaran R, Farmer A, et al. PROactive study. Lancet 2006;367:25–6, author reply 26–7.
51 Hu D, Bertozzi SM, Gakidou E, et al. The costs, benefits, and cost-effectiveness of interventions to reduce maternal morbidity and mortality in Mexico. Plos ONE 2007;2:e750.
52 Ishida H, Wong JB, Hino K, et al. Validating a Markov model of treatment for hepatitis C virus-related hepatocellular carcinoma. Methods Inf Med 2008;47:529–40.
53 Krahn M, Guasparini R, Sherman M, et al. Costs and cost-effectiveness of a universal, school-based hepatitis B vaccination program. Am J Public Health 1998;88:1638–44.
54 Kulkarni GS, Finelli A, Fleshner NE, et al. Optimal management of high-risk T1G3 bladder cancer: a decision analysis. Plos Med 2007;4:e284.
55 Lejeune C, Arveux P, Dancourt V, et al. A simulation model for evaluating the medical and economic outcomes of screening strategies for colorectal cancer. Eur J Cancer Prev 2003;12:77–84.
56 Moeremans K, Caekelbergh K, Annemans L. Cost-effectiveness analysis of bicalutamide (Casodex) for adjuvant treatment of early prostate cancer. Value Health 2004;7:472–81.
57 Niessen LW, Dijkstra R, Hutubessy R, et al. Lifetime health effects and costs of diabetes treatment. Neth J Med 2003;61:55–64.
58 Oostenbrink JB, Rutten-van Molken MP, Monz BU, et al. Probabilistic Markov model to assess the cost-effectiveness of bronchodilator therapy in COPD patients in different countries. Value Health 2005;8:32–46.
59 Ortegon MM, Redekop WK, Niessen LW. Cost-effectiveness of prevention and treatment of the diabetic foot: a Markov analysis. Diabetes Care 2004;27:901–7.
60 Perreault S, Levinton C, Laurier C, et al. Validation of a decision model for preventive pharmacological strategies in postmenopausal women. Eur J Epidemiol 2005;20:89–101.
61 Saab S, Ly D, Han SB, et al. Is it cost-effective to treat recurrent hepatitis C infection in orthotopic liver transplantation patients? Liver Transpl 2002;8:449–57.
62 Sagmeister M, Mullhaupt B, Kadry Z, et al. Cost-effectiveness of cadaveric and living-donor liver transplantation. Transplantation 2002;73:616–22.
63 Schau B, Boysen G, Truelsen T, et al. Development and validation of a model to estimate stroke incidence in a population. J Stroke Cerebrovasc Dis 2003;12:22–8.
64 Sendi PP, Craig BA, Pfluger D, et al. Systematic validation of disease models for pharmacoeconomic evaluations. Swiss HIV Cohort Study. J Eval Clin Pract 1999;5:283–95.
65 Shwartz M. An analysis of the benefits of serial screening for breast cancer based upon a mathematical model of the disease. Cancer 1978;41:1550–64.
66 Siebert U, Sroczynski G, Hillemanns P, et al. The German cervical cancer screening model: development and validation of a decision-analytic model for cervical cancer screening in Germany. Eur J Public Health 2006;16:185–92.
67 Skedgel C, Rayson D, Dewar R, et al. Cost-utility of adjuvant hormone therapies for breast cancer in post-menopausal women: sequential tamoxifen-exemestane and upfront anastrozole. Breast Cancer Res Treat 2007;101:325–33.
68 Smolen HJ, Cohen DJ, Samsa GP, et al. Development, validation, and application of a microsimulation model to predict stroke and mortality in medically managed asymptomatic patients with significant carotid artery stenosis. Value Health 2007;10:489–97.
69 Stahl JE, Vacanti JP, Gazelle S. Assessing emerging technologies—the case of organ replacement technologies: volume, durability, cost. Int J Technol Assess Health Care 2007;23:331–6.
70 Thompson KM, Segui-Gomez M, Graham JD. Validating benefit and cost estimates: the case of airbag regulation. Risk Anal 2002;22:803–11.
71 Tosteson AN, Jonsson B, Grima DT, et al. Challenges for model-based economic evaluations of postmenopausal osteoporosis interventions. Osteoporos Int 2001;12:849–57.
72 Unal B, Critchley JA, Capewell S. Explaining the decline in coronary heart disease mortality in England and Wales between 1981 and 2000. Circulation 2004;109:1101–7.
73 Urban N, Drescher C, Etzioni R, et al. Use of a stochastic simulation model to identify an efficient protocol for ovarian cancer screening. Control Clin Trials 1997;18:251–70.
74 Van Laere K, Everaert L, Annemans L, et al. The cost effectiveness of 123I-FP-CIT SPECT imaging in patients with an uncertain clinical diagnosis of parkinsonism. Eur J Nucl Med Mol Imaging 2008;35:1367–76.
75 Welsing PM, Severens JL, Hartman M, et al. The initial validation of a Markov model for the economic evaluation of (new) treatments for rheumatoid arthritis. Pharmacoeconomics 2006;24: 1011–20.
76 Munoz N, Castellsague X, de Gonzalez AB, et al. Chapter 1: HPV in the etiology of human cancer. Vaccine 2006;24(Suppl. 3):S3/1–10.

77 Aareleid T, Pukkala E, Thomson H, et al. Cervical cancer incidence and mortality trends in Finland and Estonia: a screened vs. an unscreened population. Eur J Cancer 1993;29A:745–9.
78 Devesa SS, Silverman DT, Young JL, Jr, et al. Cancer incidence and mortality trends among whites in the United States, 1947–84. J Natl Cancer Inst 1987;79:701–70.
79 Macgregor JE, Campbell MK, Mann EM, et al. Screening for cervical intraepithelial neoplasia in north east Scotland shows fall in incidence and mortality from invasive cancer with concomitant rise in preinvasive disease. BMJ 1994;308:1407–11.
80 Mahlck CG, Jonsson H, Lenner P. Pap smear screening and changes in cervical cancer mortality in Sweden. Int J Gynaecol Obstet 1994;44:267–72.
81 Cuzick J, Mayrand MH, Ronco G, et al. Chapter 10: new dimensions in cervical cancer screening. Vaccine 2006;24(Suppl. 3):S3/90–7.
82 Inglis S, Shaw A, Koenig S. Chapter 11: HPV vaccines: commercial research & development. Vaccine 2006;24(Suppl. 3):S3/99–105.
83 McCredie MR, Sharples KJ, Paul C, et al. Natural history of cervical neoplasia and risk of invasive cancer in women with cervical intraepithelial neoplasia 3: a retrospective cohort study. Lancet Oncol 2008;9:425–34.
84 Statistic New Zealand. New Zealand cohort life tables. 2009. Available from: http://www.stats.govt.nz/datasets/population/cohort-life-tables.htm [Accessed August 1, 2009].