

Number XX

Decision and Simulation Modeling: Review of Existing Guidance, Future Research Needs, and Validity Assessment

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

This information does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services determination or policy.

Contract No.

Prepared by:

Investigators:

AHRQ Publication No.

<Month Year>

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted, for which further reproduction is prohibited without the specific permission of copyright holders.

Suggested Citation: pending

None of the investigators has any affiliations or financial involvement that conflicts with the material presented in this report.

This report is based on research conducted by an Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. xxx-xxxx-xxxxx). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

This report may be used, in whole or in part, as the basis for the development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Richard Kronick, Ph.D.
Director
Agency for Healthcare Research and
Quality

Yen-pin Chiang, Ph.D.
Acting Deputy Director, Center for
Evidence and Practice Improvement
Agency for Healthcare Research and
Quality

Stephanie Chang, M.D., M.P.H.
Director, EPC Program
Center for Evidence and Practice
Improvement
Agency for Healthcare Research and
Quality

William Lawrence, M.D., M.S.
Task Order Officer
Center for Evidence and Practice
Improvement
Agency for Healthcare Research and
Quality

Contents

Introduction	1
Chapter 1. Systematic Review of Recommendations For the Conduct and Reporting of Decision and Simulation	
Models.....	2
Introduction	2
Methods.....	3
Systematic Review.....	3
Stakeholder Meeting.....	4
Results.....	5
Systematic Review of Published Recommendations.....	5
Stakeholder Panel Discussions	9
Conclusions	9
Chapter 2. Review of Guidance from Health Technology Assessment Organizations	11
Introduction	11
Methods	11
Results.....	12
Integration of modeling.....	16
Modeling alongside systematic review	16
Timing of modeling.....	16
Use of pre-existing or established models	16
Modeling recommendations	17
How systematic reviews were incorporated into the modeling process	17
Who conducts the model	17
Inclusion of quality of life	17
Inclusion of costs	17
Budget analysis done.....	17
Impact on project budget.....	17
Conclusions	17
Chapter 3. Future Research Needs For Decision and Simulation Modeling	20
Introduction	20
Methods.....	20
Results.....	21
Conclusions	27
Chapter 4. A review of validation and calibration methods for healthcare decision and simulation models	29
Introduction	29
Methods.....	29
Sources of information and review methods.....	29
Definitions and preliminaries	30
Results: Overview of Validation and Calibration Methods	31
Model validation	31
Model calibration	33
Methodological appraisals of validation and calibration methods in health-care models	35
Calibration as estimation.....	37
Conclusions	39
References	41

Tables

Table 1: Recommendation types, by process of generation.....	8
Table 2. Proposed questions for categorization of HTA guidance	12
Table 3. Summary of available guidance on modeling from international HTA agencies	13
Table 4. Dimensions of need (Approach to prioritization)*	21
Table 5. List of future research needs derived from the systematically reviewed articles.....	23
Table 6. List of future research needs derived from stakeholders	24
Table 7. Studies comparing alternative calibration methods applied to the same problem.....	36

Figures

Figure 1. Flow diagram for papers presenting recommendations for decision and simulation modeling	6
Figure 2. Map of recommendation statements from the systematic literature review.	7
Figure 3. Diagram of the modeling process	31

Appendices

Appendix A	Acknowledgments
Appendix B	Systematic review search strategy
Appendix C	Health Technology Assessment Organizations
Appendix D	HTA Data Extraction

Appendices cited in this report are available at <http://ahrq.gov/XXX> [AHRQ will supply the rest of the information for this statement.]

Abstract

Background. Despite rigorous systematic reviews of efficacy and effectiveness of healthcare interventions, patients, providers and policymakers may remain in doubt about what they should do because of uncertainty, tradeoffs among benefits and harms, and conflicting preferences. Decision and simulation models can supplement systematic reviews to increase the usefulness of the evidence summary. The aims of this report are four-fold: 1) to summarize evidence- and consensus-based guidance on the conduct and reporting of decision and simulation models; 2) to summarize guidance from Health Technology Assessment (HTA) groups for modeling; 3) to prioritize future research needs to improve models; and 4) to provide an overview of methods for model calibration and validation.

Methods. With guidance from a Technical Expert Panel and a Clinical and Policy Advisory team with clinical, methodological research, and policymaking expertise, we completed five projects. For Aim 1, we conducted a systematic review of articles that provided evidence- or consensus-based recommendations for the conduct and reporting of decision or simulation models. We classified recommendation statements into four domains: model structure, data, consistency, and communication of model results. To contextualize the findings of the systematic review, we organized a meeting with a group of 28 stakeholders, including modelers, users of models, and funders of research. For Aim 2, we searched the websites of 126 international agencies and institutes conducting HTA for real-world practices regarding when to apply decision and simulation modeling methods, which we summarized. For Aim 3, from the systematic review and from the stakeholders in Aim 1 we identified and collected suggestions for future research needs. Stakeholders prioritized those needs based on importance, desirability of new research, feasibility, and potential impact. For Aim 4, we searched for articles that compared or applied alternative validation methods for decision or simulation models. We extracted and summarized descriptions and comparisons of any methods and reported results for face validity, internal validity, external validity, cross-model validation, and calibration.

Results. The systematic review of modeling recommendations found 71 eligible articles. 90% of articles (n=64) provided recommendations regarding model structure. Almost all articles (n=68, 96%) also provided recommendations regarding obtaining appropriate data to populate models. Stakeholders highlighted the importance of establishing guiding principles for “good practice” but discouraged the use of “cookbook” checklists. Of the 71 articles, 38 (54%) provided suggestions for future research; stakeholders provided 28 additional suggestions. 21 HTA organizations provided guidance through their websites regarding the application of decision modeling in the context of conducting a HTA. The prioritized future research needs included questions about model data, model structure, consistency, and reporting. Reviewed studies provided information on model validation (face validity, verification and internal validation, external validation, and cross-model validation) and calibration (varying specifications of the calibration problem with the same and different algorithms and use of alternative algorithms for the same calibration problem).

Conclusion. Our systematic review and stakeholder meeting summary provides a comprehensive compendium of guidance documents for decision and simulation modeling, annotated with

information on the domains covered by each document, and the methods used to arrive at specific recommendations. We also summarized recommendations for conducting models for HTA organizations. These processes enabled us to prioritize future research needs to form an empirical basis for and to improve recommendations for modeling. Our overview of model validation and calibration provides insights into the relative value and efficiency of different methods.

Introduction

Despite rigorous systematic reviews of healthcare interventions, patients, providers and policymakers may remain in doubt about which interventions to use because of uncertainty, tradeoffs between benefits and harms, and differences in preferences. First, even after a synthesis of best-available evidence, uncertainty may remain because the studies have not adequately addressed patient-relevant outcomes or the applicability of studies to patients may be poor. Second, tradeoffs between benefits and harms occur. For example, the U.S. Preventive Services Task Force (USPSTF) analysis of mammography for women in their 40s suggests a statistically significant reduction in breast cancer death but also potential harms due to radiation exposure, overdiagnosis and overtreatment.¹ Thus, optimal decisionmaking for individuals and populations may depend on their values (or preferences) regarding the different potential outcomes. Lastly, the choice between different interventions usually results in different resources being used. Just as the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group considers resource utilization in guideline development, the USPSTF modeled the effect of different intervention choices to estimate resource consumption for its recommendations on cancer screening. The models allowed fuller consideration of the results and implications for patients and society of choosing healthcare intervention alternatives.

This project was designed to advance the credibility, transparency, and methodological rigor of modeling performed alongside systematic reviews by 1) assembling a multidisciplinary modeling core methods group; 2) extending current Evidence-based Practice Center (EPC) guidance on developing models alongside systematic reviews by reviewing and evaluating current literature on best practices in prioritization, conduct, and dissemination of modeling; 3) identifying priority topics in methods research for decision and simulation modeling; and 4) conducting a pilot methods study.

The report is written as a series of five stand-alone chapters, each in a manuscript format. Chapter 1 is a systematic review of recommendations for the conduct and reporting of decision and simulation models. Chapter 2 is a review of guidance provided by international Health Technology Assessment (HTA) websites. Chapter 3 prioritizes future research needs to improve the conduct and reporting of decision and simulation models. Chapter 4 describes a study to address a prioritized future research need. In this chapter, we review reports of models that used multiple methods to either validate or calibrate their models.

Chapter 1. Systematic Review of Recommendations For the Conduct and Reporting of Decision and Simulation Models

Introduction

Rigorous systematic reviews of the literature have become the preferred way to identify, appraise, and synthesize studies on the comparative effectiveness of competing healthcare interventions. However, users of such reviews—including patients, providers, and policymakers—may remain in doubt about which interventions to use because of persistent uncertainties, tradeoffs between benefits and harms, differences in preferences, or insufficient evidence.^{2,217} For instance, even after a synthesis of best-available evidence, uncertainty may remain regarding the optimal choice among available interventions for important patient-relevant outcomes because studies may provide information mainly on surrogate outcomes, have short follow-up durations, or report inadequate subgroup data.

Policy needs often require decision making under uncertainty and decision makers have become increasingly interested in complementing the results of systematic reviews of empirical evidence with information from decision and simulation modeling. Modeling is especially useful when uncertainty exists about how specific evidence might be applied to a particular decisional context or a specific patient, or formulated into a recommendation. Modeling can provide a comprehensive, transparent, and interpretable integration of empirical evidence on benefits and harms, values (preferences), and resource utilization, while accounting for all relevant sources of uncertainty.³ Model results can be used to guide decision making and to support the prioritization and planning of future clinical research activities.^{4;5} Decision-making typically involves trade-offs in harms and benefits and thus may depend on individuals' values for a range of outcomes. Modeling offers a coherent approach for integrating clinical research evidence and patient values to optimize decision making.

Methods for the conduct of decision and simulation modeling have continued evolving to address the ever-increasing information needs of decision makers. The complexity and continued advances of the relevant methods have spurred the publication of recommendation statements on “best practices” for decision and simulation modeling. Two previous systematic reviews have assessed published recommendations and guidelines for decision analytic modeling for health technology assessment in an effort to identify methodological recommendations for decision-analytic modeling. Philips et al., in 2004, identified existing guidelines for best practices in health technology assessment and classified them into three domains—model structure, data, and consistency.^{6,216} Kuntz et al., in 2013, reviewed recommendations published through 2009 for developing, validating and using decision-analytic models in the context of systematic reviews.^{2,217} In addition to model structure, data, and consistency, they assessed a fourth domain pertaining to the communication of model results.

We sought to update and expand previous syntheses of evidence- and consensus-based guidance on modeling by performing a systematic review of published recommendation statements. To help contextualize and interpret the systematic review findings, we convened a meeting of diverse stakeholders, including modelers, users of models, and funders of research.

Here we summarize the results of the systematic review and stakeholder input, to provide an up-to-date compendium of recommendations for best practices in decision and simulation modeling.

Methods

The Tufts Evidence-based Practice Center, under contract with the Agency for Healthcare Research and Quality (AHRQ), conducted a systematic review of the published literature for evidence- and consensus-based guidance on the conduct of decision and simulation models. We convened a Clinical and Policy Advisory team (CaPA), and a Technical Expert Panel (TEP) to provide input in the design and conduct of the review (see **Appendix A**). The CaPA was formed to provide clinical and policymaking expertise from individuals who have used modeling input to inform decision-making. It included three members, with expertise and experience in applying decision analysis and simulation modeling for developing clinical guidelines, assessing public health risk management, and informing healthcare policy decisions. The TEP included four internationally-recognized experts in evidence synthesis and decision modeling and two policymakers and payers, who also have experience in decision modeling.

Systematic Review

Search strategy

We searched four electronic databases (MEDLINE, Cochrane Methodology Register, Health Technology Assessment Database, NHS Economic Evaluation Database) for articles presenting best practices in prioritization, conduct, and dissemination of decision and simulation modeling through October 30, 2012. We based our search strategy on that used by Philips et al. and Kuntz et al. but substantially expanded the search terms to include keywords related to modeling methods and guideline statements.^{2;6;216;217} We confirmed that our search captured all articles reviewed by the two prior reviews, relevant articles from our personal bibliographies, and those suggested by CaPA, and TEP members. To make the size of the corpus that needed to be screened manageable (the search yielded 65,053 citations), we limited our screening to 37 journals identified with input from the CaPA and TEP as likely to publish guidance documents pertaining to decision analysis, computer simulation, health economic analysis for technology assessment, and health outcomes research. We also limited the search to articles published since 1990. The final search strategy, with the list of included journals, is presented in **Appendix B**.

Abstract screening and study selection

Using the open-source abstract screening software Abstrackr (<http://www.cebm.brown.edu/software>),⁷ five reviewers independently screened abstracts in duplicate and resolved disagreements by group consensus. Eligible studies had to provide guidance on the elements of a good decision-analytic model, address the key elements that constitute a good decision analytical or simulation model, or provide explicit criteria against which to assess the quality or validity of a decision-analytic model. Furthermore, we required that the guidance had to be either evidence-based (i.e., from a systematic review) or consensus-based (e.g., from discussion or collaboration of experts) and that articles provided a description of the process through which the authors arrived at their recommendations. Although this requirement excluded some older seminal articles,⁸⁻¹¹ we found that their recommendations had been incorporated into more recent eligible articles. Examples of acceptable processes included systematic reviews, a Delphi consensus process, or presentation at

a conference with active feedback from meeting attendees. We excluded articles that appeared to be based only on the opinions of the authors.

Eligible recommendation statements had to provide general guidance for decision and simulation modeling (e.g., guidance applicable across multiple disease topics such as comparative effectiveness of treatments, screening strategies, infection control, telemedicine) or for making formulary decisions at a national or regional level. Because of the potential for limited generalizability, we excluded guidance for modeling of specific diseases (e.g., cancer), local decision makers (e.g., hospital formularies), single modeling methodologies (e.g., Markov model, microsimulation model), or specific aspects of the modeling process (e.g., uncertainty propagation in models, reporting or dissemination of modeling guidance or results). This was done to constrain the scope of the project and because guidance on specific model aspects is incorporated in documents providing more broad guidance (e.g., the recent ISPOR-SMDM guidance covers all key specific aspects of modeling⁷⁸). Thus, the guidance had to provide recommendations for a range of methodologies for and components of decision-analytic modeling.

We further excluded articles that provided recommendations for decision aids or for statistical analyses for estimating effect size, inference, or prediction; evaluated only costs; were related to clinical chemistry and laboratory medicine, occupational health, or vaccines; or dealt with multi-criteria decision analysis and analytic hierarchy process methods.

Data Extraction

We created a data extraction form based on the three-component framework originally proposed by Sculpher et al. in 2000⁹ and employed in a previous systematic review of methodological recommendations:^{6,216} model structure, model data, and model consistency. In addition, we added a component pertaining to the “communication of model results” as suggested by Kuntz et al. which we also found to be a useful addition to the original framework.^{2,217} All subcategories within the four thematic components were identified, and an operational definition for each subcategory was determined using the checklist published by Philips et al.⁶ In each article, we identified specific recommendations as statements of policy or procedure that used modal verbs (e.g., “should,” “must”), that were placed in the context of other text as a suggestion for action, that were explicitly noted as a recommendation, or that were included in a checklist of items used to critically review a model. We then mapped each recommendation statement to the appropriate component—structure, data, consistency, presentation of results—and subcategories therein. For each publication, we also extracted the process for generating recommendations, the purpose of the recommendations or article, the target audience, and the intended scope of the recommendations.

Stakeholder Meeting

We invited workshop participants who would represent expertise in decision and simulation modeling, systematic review and evidence-based medicine, epidemiology, and biostatistics, and perspectives from six stakeholder groups: patient representatives, providers of care, purchasers of care, payers, policy makers (including research funders and professional societies), and principal investigators¹² In total 43 individuals were invited to participate and 28 attended the 1-day meeting in-person or remotely.

The goals of the workshop were to review and expand the list of recommendations and research needs identified by the systematic review of methodological recommendations for decision and simulation modeling and to develop a list of priority research areas aiming to

improve the usefulness and credibility of models used to inform decision making. Results regarding future research needs will be reported separately.

One of the authors (JBW) began the meeting with a presentation about decision modeling to ensure common background knowledge and vocabulary and to set the ground rules for the ensuing discussions. We then presented the categories and topics of recommendation statements identified by the systematic review, followed by a group discussion involving all participants. After assignment into three smaller groups (each comprising 7 to 12 participants), stakeholders reviewed and discussed the modeling recommendations in-depth over two breakout sessions. One investigator facilitated each session (JBW, EMB, and IJD) and a second investigator kept detailed notes (DM, TAT, or JAC) to supplement the tape-recording of the discussions. Facilitators reviewed the goals for each session and used a list of topics to guide the unstructured discussions. Stakeholders were encouraged to comment on available recommendations and identified gaps, limitations and areas for expansion within each component of interest. Key points from each of the small group discussions were presented to the whole group for further discussion. Following the meeting, we circulated summaries from all discussions to participants and solicited additional comments via email.

Results

Systematic Review of Published Recommendations

The search (**Appendix B**) yielded 6825 citations (**Figure 1**); the TEP and CaPA suggested 20 additional articles. We reviewed 358 articles in full text, of which 71 met our eligibility criteria and reported recommendations for conducting decision and simulation modeling.^{13-51;6;8;11;52-77,216} Of these, 29 provided recommendations based on expert panel deliberations, 14 on nonsystematic literature reviews, 14 on systematic literature reviews, 7 on conference discussions, and 8 on a combination of these methods.

The complete list of extracted information is available electronically on the Systematic Review Data repository (project title: *Recommendations for decision and simulation modeling*).

Figure 1. Flow diagram for papers presenting recommendations for decision and simulation modeling

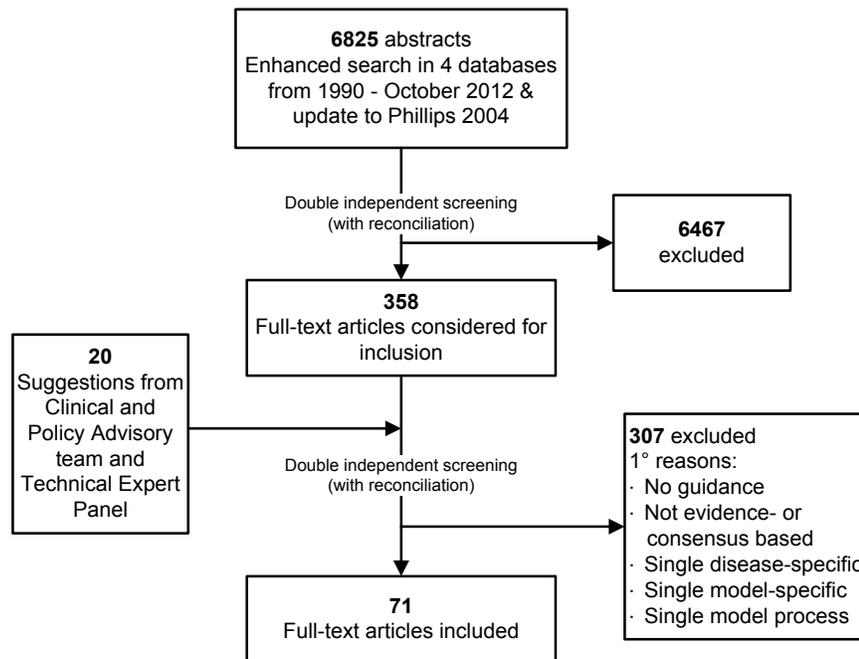


Figure 2 presents a summary of the recommendation types extracted from each article, along with information on the process used to generate the recommendations. When examining a single row of the matrix, a preponderance of filled cells indicates that most articles have provided recommendations addressing a specific area. For example, defining the scope and perspective of the analysis is the most common area of recommendations across the articles we reviewed. When examining a single column of the matrix, a preponderance of filled cells indicates that a given paper has provided recommendations addressing many different areas. For example, the paper by Phillips et al. (2004) addressed all but two of the areas we examined.⁶

The majority of articles (64/71; 90%) provided one or more recommendations regarding model structure, including modeling objectives (40/71; 56%); model scope and perspective (52/71; 73%), and choice and justification of comparators used in modeling (46/71; 65%). Nearly all (68/71; 96%) made recommendations about obtaining appropriate data to populate models (main and sensitivity analyses), including obtaining cost data (49/71; 69%), methods for data identification and synthesis (40/71; 56%), and the conduct of sensitivity analyses (40/71; 56%). A minority of articles provided recommendations regarding the internal, external, or predictive validity of models (27/71; 38%). Recommendations about model validity mainly pertained to internal (12/71; 17%) and external validity (18/71; 26%), with only a small number addressing the predictive validity of models (4/71; 6%). **Table 1** gives the count of recommendation types for all 71 articles and stratified by process used to generate the recommendations. Overall, no clear pattern of association between process and specific recommendation types was apparent.

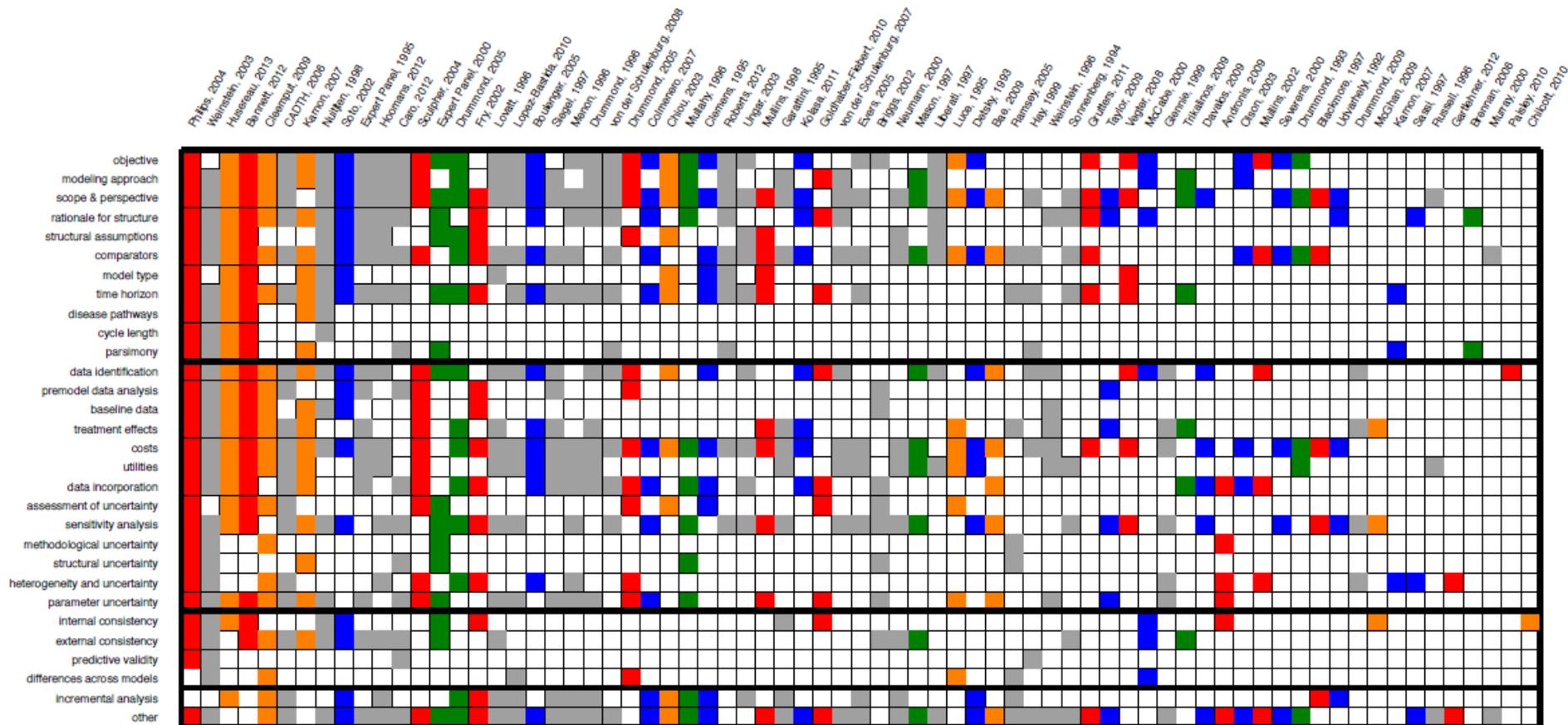


Figure 2. Map of recommendation statements from the systematic literature review.

The figure presents a visual summary of the recommendation statements extracted from the systematic review. Columns represent the individual eligible papers (sorted left to right by number of recommendation subcategories). Rows represent recommendation topics. Thick black lines group subcategories belonging to the same component (from top to bottom: structure, data, consistency, presentation of results and miscellaneous). Red boxes indicate recommendations based on systematic reviews, grey boxes indicate expert panels, green boxes indicate conferences or meeting, blue boxes indicate “nonsystematic” reviews, and orange boxes indicate combinations of methods.

Table 1: Recommendation types, by process of generation

Reporting Items, by category		All documents N=71	SR N=14	Expert panel N=28	Conf/Mtg N=7	Non-SR N=14	Combo N=8
STRUCTURE	objective	40 (56)	7 (10)	15 (21)	4 (6)	9 (13)	5 (7)
	modeling approach	32 (45)	5 (7)	15 (21)	4 (6)	4 (6)	4 (6)
	scope & perspective	52 (73)	9 (13)	22 (31)	6 (8)	10 (14)	5 (7)
	rationale for structure	33 (46)	5 (7)	14 (20)	3 (4)	8 (11)	3 (4)
	structural assumptions	17 (24)	5 (7)	7 (10)	2 (3)	1 (1)	2 (3)
	comparators	46 (65)	8 (11)	22 (31)	3 (4)	8 (11)	5 (7)
	model type	12 (17)	4 (6)	3 (4)	0 (0)	2 (3)	3 (4)
	time horizon	36 (51)	7 (10)	17 (24)	3 (4)	5 (7)	4 (6)
	disease pathways	6 (8)	2 (3)	2 (3)	0 (0)	0 (0)	2 (3)
	cycle length	5 (7)	2 (3)	2 (3)	0 (0)	0 (0)	1 (1)
	parsimony	12 (17)	2 (3)	5 (7)	2 (3)	1 (1)	2 (3)
	DATA	data identification	40 (56)	8 (11)	17 (24)	3 (4)	7 (10)
premodel data analysis		15 (21)	5 (7)	6 (8)	0 (0)	2 (3)	2 (3)
baseline data		12 (17)	4 (6)	4 (6)	0 (0)	1 (1)	3 (4)
treatment effects		25 (35)	4 (6)	11 (15)	2 (3)	3 (4)	5 (7)
costs		49 (69)	9 (13)	20 (28)	4 (6)	10 (14)	6 (8)
utilities		28 (39)	3 (4)	17 (24)	2 (3)	2 (3)	4 (6)
data incorporation		30 (42)	8 (11)	10 (14)	3 (4)	6 (8)	3 (4)
assessment of uncertainty		13 (18)	5 (7)	2 (3)	1 (1)	1 (1)	4 (6)
sensitivity analysis		40 (56)	6 (8)	19 (27)	4 (6)	7 (10)	4 (6)
methodological uncertainty		6 (8)	2 (3)	2 (3)	1 (1)	0 (0)	1 (1)
structural uncertainty		8 (11)	1 (1)	4 (6)	2 (3)	0 (0)	1 (1)
heterogeneity and uncertainty		18 (25)	7 (10)	6 (8)	1 (1)	3 (4)	1 (1)
parameter uncertainty		29 (41)	7 (10)	13 (18)	2 (3)	2 (3)	5 (7)
CONSISTENCY	internal validity	14 (20)	5 (7)	3 (4)	1 (1)	2 (3)	3 (4)
	external validity	19 (27)	2 (3)	10 (14)	3 (4)	2 (3)	2 (3)
	predictive validity	4 (6)	1 (1)	3 (4)	0 (0)	0 (0)	0 (0)
	differences across models	7 (10)	1 (1)	3 (4)	0 (0)	1 (1)	2 (3)
OTHER	incremental analysis	24 (34)	2 (3)	12 (17)	2 (3)	5 (7)	3 (4)
	other, including reporting	52 (73)	9 (13)	24 (34)	5 (7)	11 (15)	3 (4)

All results are expressed as number of papers (percentage of papers, out of 71).

Combo = combination of methods, Conf/Mtg = conference or meeting, SR = systematic review.

Stakeholder Panel Discussions

The stakeholder group, together with the CaPA, TEP, and the investigators, as well as representatives from AHRQ, met for a day-long series of sessions on 27 February 2013 at AHRQ headquarters (Rockville, MD), to comment and deliberate on the findings of the systematic review (see **Appendix A**). After reviewing the list of recommendations, stakeholders uniformly agreed on the need to increase transparency of various modeling processes. Initial discussions centered on clarifying individual recommendations and, occasionally, their classification into the four components (structure-data-consistency-communication of results). Stakeholders highlighted the importance of recommendations that apply broadly, across multiple types of models. Questions were raised on the appropriateness of using the Philips et al. report as a framework and whether individual recommendations should be bundled/re-bundled or stand alone.^{6,216} While recognizing that the classification of recommendations into the four components involved judgments (albeit our best), we believe that it will be helpful to modelers seeking to identify guidance related to a specific aspect of their work.

In small group discussions, stakeholders questioned the value of yet another “checklist of recommendations”. Many felt it would be more helpful to identify “best practices” or “principles” on how to integrate modeling and the systematic review processes. They stressed the importance of understanding the needs and perspectives of the different stakeholders who might use model results, with particular consideration of patient preferences. Participants strongly indicated that the presentation and contextualization of modeling results to different stakeholders (e.g., policy-makers) remains an understudied area that is as important as developing methodological guidance for the conduct of decision modeling and simulation studies.

Conclusions

This systematic review updates and organizes recommendations for decision and simulation modeling in the medical literature into four domains. It classifies recommendations by domain and by the methods used to develop them. One might conjecture that the recommendation topics that were most frequently addressed in the papers we reviewed were either the most important (as perceived by those making the recommendations) or the easiest to address (on the basis of available theoretical or empirical evidence). Over 90% of recommendation documents provided one or more recommendations pertaining to model structure and data appropriateness, but only a minority provided recommendations regarding model validity. Given the critical importance of assessing model validity,⁷⁸ the absence of recommendations on methods for model validation more likely reflects uncertainty about the most appropriate methodological approaches.

We discussed the findings of the systematic review with a diverse stakeholder panel that included patient representatives, providers of care, purchasers of care, payers, policy makers and principal investigators, many with expertise in decision and simulation modeling, systematic review methods, epidemiology, and biostatistics. These discussions highlighted the importance of transparency in modeling methods in decision analysis applications and the need to identify best practices, rather than use “cookbook” checklists. When applied to complex methodological decisions (such as the design and conduct of decision and simulation modeling studies) checklists tend to promote mechanistic approaches that do not adequately address the practical challenges faced by modelers. Alternatively, determining the theoretical considerations relevant to each research problem and working from “first principles” may result in more fundamentally

sound models. Nonetheless, checklists can increase consistency in the reporting of decision models and their findings, and in the peer review of modeling research,^{31;79}

Our work has limitations that need to be considered when interpreting our results. The majority of the recommendation statements we reviewed were derived from the medical literature, leading to a preponderance of statements regarding economic analyses (using decision and simulation models). Nonetheless, outside of cost-related considerations, many of the recommended methodological principles apply across multiple decision and simulation model types, as presented in our annotated bibliography of recommendations. More generally, systematic reviews of methodological topics are usually less likely to be comprehensive than reviews of empirical research studies (e.g., clinical trials) because of the unavailability of standardized indexing terms and the large number of sources that need to be searched. Further, we only classify guideline statements into four broad domains (components) and sub-components, without attempting to reconcile conflicting statements or to provide a synthesis across statements. Such a synthesis would be an appropriate task for a recommendation-making panel, such as the group updating the recommendations of the Panel on Cost Effectiveness in Health and Medicine (personal communication, Prof. Peter Neumann).⁸⁰

In conclusion, this systematic review and stakeholder meeting summary provides a comprehensive compendium of guidance documents for decision and simulation modeling, annotated with information on the domains covered by each document, and the methods used to arrive at specific recommendations. Our annotated bibliography will be useful to modelers and others looking for sources of evidence- or consensus-based guidance. We have also incorporated this review of the existing guidance into a separate document proving *Guidance for the Conduct and Reporting of Decision and Simulation Models in the Context of Health Technology Assessment*. [Reference forthcoming]

Chapter 2. Review of Guidance from Health Technology Assessment Organizations

Introduction

Health technology assessment (HTA) is a method of evidence synthesis that utilizes systematic review methodology to study the medical, social, ethical, and economic implications of development, diffusion, and use of health technology.⁸¹ HTA is evolving with the need for continued decisionmaker support when evaluating new technologies.⁸² Systematic review, utilizing meta-analytic techniques, provides best estimates of average effects of interventions and rates of outcomes (including benefits and harms). Given sufficient evidence, systematic review may also provide an indication about how effects of interventions may differ in different settings, among different people, and across variations of the interventions (e.g., different doses or diagnostic thresholds). However, systematic reviews alone often provide a poor basis to balance benefits and harms in different scenarios or for different patients, to balance benefits and the resources required to achieve them, or to incorporate individual values (preferences) into decisionmaking. Decision modeling can provide a comprehensive, transparent, and interpretable integration of empirical evidence on benefits and harms, preferences, and resource utilization, while accounting for all relevant sources of uncertainty. However, it is hard to determine when a decision model may be of sufficient value to justify the time and resources required to incorporate it alongside systematic reviews.⁸³

The goal of this review is to identify and summarize guidance from international HTA organizations, agencies and institutes that assess new health technologies or economic evaluations regarding when and how to incorporate decision modeling alongside systematic reviews, specifically searching for descriptions regarding whether to integrate modeling as part of HTA, the timing of modeling in relation to the conduct of systematic review, modeling methodological recommendations, the potential budget impact of introducing the technology, and the effect of including modeling on the HTA project budget. We then categorized all available recommendations across HTA organizations to examine both differences and commonalities in guidance.

Methods

Mathes et al. originally identified the HTA organizations through member lists of the International Network of Agencies for Health Technology Assessment (INAHTA), Health Technology Assessment International (HTAi) not-for-profit organizations, and the European Network for Health Technology Assessment (EUnetHTA).⁸⁴ Building on this review we searched the websites of 126 international agencies and institutes conducting HTA (HTA organizations) for guidance regarding when and how to apply decision modeling methodology. For our review, two researchers independently reviewed identified HTA organization websites in January 2014 (**Appendix C**). Standardized screening procedures included use of website navigation (e.g., hyperlinks utilizing key terms such as "publications", "recommendation" and "methods"), website search engines, and sitemaps to identify appropriate Web content and documents.

Relevant data were collected verbatim directly from website text and linked documents available through the website such as handbooks or guidelines on HTA methods. Non-English language websites and documents were excluded from extraction. From each source, we extracted language regarding when and how to model including integration of modeling as part of HTA, modeling alongside systematic review of literature, timing of modeling with respect to the systematic review (i.e., concurrent or sequential), use of pre-existing versus established models, how to model, and budgetary considerations. Relevant language was extracted verbatim to standardized spreadsheets. Reviewers further categorized extracted language on when to model into 11 categories, as listed in **Table 2**. All extracted language and categorizations were reviewed by the entire project team for consensus.

Table 2. Questions for categorization of HTA guidance

Category	Question
Integration of modeling	Is modeling always a part of HTA?
Modeling alongside systematic review	Is modeling always performed alongside a systematic review?
Timing of modeling	Should modeling be performed concurrent with or after a systematic review once the parameters are set?
Use of pre-existing or established models	Should one use pre-existing or established models? Include the process for this decision (e.g., Is a systematic review of models conducted?; How is a model deemed adequate for use?)
Modeling recommendations	What are the methods for modeling?
How systematic review incorporated into the model	How should one incorporate the systematic review into the model?
Who conducts the model	Who should perform the modeling?
Inclusion of quality of life	Should the model include quality of life?
Inclusion of costs	Should the model include costs?
Budget analysis done	Should one conduct a budget analysis (e.g., a national budget impact analysis)?
Impact on project budget	Should one conduct an analysis of the impact of modeling on a project budget?

Results

Of the 126 websites, 21 (17%) provided relevant text regarding the application of decision modeling in the context of conducting a HTA.⁸⁵⁻¹⁰⁵ The remaining 106 websites (84%) did not provide any guidance or were written in non-English languages and could not be translated (42 websites, 33%). The 21 websites with relevant, extractable data included HTA organizations from four continents including two HTA organizations in Asia (Taiwan and Thailand), three in Australia and New Zealand, 11 in Europe, and five in North America (3 U.S. and 2 Canada). A summary of available guidance on modeling across the 21 websites is presented in **Table 3 and Appendix D**. Details specific to each of the 11 subcategories are described below.

Table 3. Summary of available guidance on modeling from international HTA agencies

Website	Integration of modeling	Modeling alongside SR	Timing of modeling	Use of pre-existing or established models	Modeling recommendations	How SR incorporated into the model	Who conducts the model	Inclusion of quality of life	Inclusion of costs	Budget analysis done	Impact on project budget
Agency for Health Technology Assessment in Poland (AHTApol/Poland)	X				X			X	X		
Canadian Agency for Drugs and Technologies in Health (CADTH/Canada)					X				X		
Danish Centre for Health Technology Assessment (DACEHTA/Denmark)	X				X			X	X		
Health Information and Quality Authority (HIQA/Ireland)	X	X	X		X			X	X	X	
National Authority of Medicines and Health Products (INFARMED/Portugal)			X		X						
Institute for Quality and Efficiency in Health Care (IQWiG/Germany)	X								X		
Belgian Federal Health Care Knowledge Centre (KCE/Belgium)	X							X	X		
MAS (Medical Advisory Secretariat, within the Ontario Ministry of Health and Long-Term Care Health Strategies Division/Canada)					X				X		

Website	Integration of modeling	Modeling alongside SR	Timing of modeling	Use of pre-existing or established models	Modeling recommendations	How SR incorporated into the model	Who conducts the model	Inclusion of quality of life	Inclusion of costs	Budget analysis done	Impact on project budget
Medical Services Advisory Committee (MASC/Australia)					X				X		
National Institute for Clinical Excellence (NICE/UK)	X	X		X	X	X		X	X	X	
Pharmaceutical Benefits Advisor Committee (PBAC, Australia)	X				X						
Pharmaceutical Management Agency of New Zealand (PHARMAC/New Zealand)	X										
AAZ (Agency for Quality and Accreditation in Health Care/Croatia)					X				X		
HITAP (Health Intervention and Technology Assessment Program/Thailand)	X				X				X	X	
ICER (Institute for Clinical and Economic Review/US)	X	X		X	X	X					
LBI (Ludwig Boltzmann Institute for Health Technology Assessment/Austria)	X				X						
MHRA (Medicines and Healthcare Products Regulatory Agency/UK)	X				X						

Website	Integration of modeling	Modeling along-side SR	Timing of modeling	Use of pre-existing or established models	Modeling recommendations	How SR incorporated into the model	Who conducts the model	Inclusion of quality of life	Inclusion of costs	Budget analysis done	Impact on project budget
NLM (National Library of Medicine/US)	X				X			X	X		
AHRQ (US Agency for Healthcare Research and Quality/US)	X		X	X	X	X	X	X	X		X
CAST (Centre for Applied Health Services Research and Technology Assessment, University of Southern Denmark)	X	X	X		X				X		
CDE (Center for Drug Evaluation/Taiwan)		X								X	

Abbreviation: SR, systematic review

Despite the relatively large number of international HTA organization websites, only 17% (21 agencies) provided guidance regarding the application of decision modeling in the context of conducting an HTA. On average, each website provided guidance on roughly 4 of the 11 subcategories regarding when and how to incorporate decision modeling. AHRQ provided the most guidance among the 21 websites, addressing 9 out of 11 subcategories, followed by NICE and Health Information and Quality Authority (HIQA/Ireland) providing guidance on 8 and 7 of the subcategories, respectively. All but one HTA organization (Pharmaceutical Management Agency of New Zealand) provided guidance on multiple subcategories. The most frequently reported guidance across HTA organization websites addressed recommendations on how to model (17 agencies) and integration of modeling as part of HTA (15 agencies); only one HTA organization (AHRQ) provided guidance on who develops and implements the model.

Integration of modeling

Fifteen HTA organizations addressed incorporation of modeling in the context of conducting a HTA. Of these 15, six agencies required a decision model; three as part of HTA development process and three as part of economic evaluation process. Three other agencies recommended, but did not require, the incorporation of the decision model as part of HTA under certain conditions. For example, two recommended incorporating decision models when conducting an economic evaluation. The remaining six agencies neither required nor recommended modeling as part of HTA. Rather they remained neutral on the topic while acknowledging that the technique has been used in prior agency reports.

Modeling alongside systematic review

Five HTA organizations addressed modeling alongside systematic review. One mentioned modeling but did not specify whether it should always be incorporated alongside systematic review or alongside HTA in general. Of the remaining four HTA organizations, only one website stated that a decision model must always accompany a systematic review when conducting HTA. The other three did not always require a model.

Timing of modeling

Four HTA organizations provided guidance on the timing of modeling in the context of conducting a HTA. Two recommended that the modeling be done concurrently with the systematic review while the other two recommended the modeling be done after the review once parameters have been estimated with one, however, acknowledging that it may not be feasible or timely to conduct the systematic review and model in sequence rather than in parallel.

Use of pre-existing or established models

Only three of the 15 HTA organizations addressed the use of pre-existing versus established models in the context of conducting a HTA. One did not favor the use of pre-existing or established models in HTA and recommended the development of a *de novo* model to accompany each systematic review. The remaining two HTA organizations acknowledged that it was acceptable to use pre-existing models under certain conditions; one required that the model be conducted by manufacturers and sponsors of the HTA and the other cautioned against using established models that are not flexible enough to represent the consequences of all interventions of interest.

Modeling recommendations

Seventeen of the 21 HTA organizations (81%) provided guidance on how to model. We divided these recommendations into four categories addressing whether or not the advice featured a statement on data, structure, validity and assumptions. Twelve of the 17 HTA organizations included a statement on model data, 14 addressed model structure, 6 addressed model validation, and 7 addressed model assumptions.

How systematic reviews were incorporated into the modeling process

Whether a model is prepared concurrently with or after the completion of a systematic reviews, the question remains how to incorporate the review results into the model. Three HTA organizations gave guidance on this issue with two stating that the model outcome estimates should be based on the systematic reviews. The remaining HTA organizations suggested that the systematic reviews be used to produce parameter estimates for use in sensitivity analyses.

Who conducts the model

AHRQ was the only HTA organization addressing the question of who specifically should perform the modeling. They note that it is not always feasible for the systematic review team to also conduct a model since different expertise is needed. When separate teams are used, they should collaborate closely and should, ideally, reside in the same location.

Inclusion of quality of life

All seven HTA organizations that comment on quality of life suggest that models should take into account differences in quality of life among health states. Four of the seven state that a model should always include quality of life. The other three state that quality of life should be incorporated when appropriate (e.g., when final utility results are needed or when there is adequate evidence about quality of life to include in the model).

Inclusion of costs

Fourteen HTA organizations provided advice on whether costs should be included in a model. Only four HTA organizations stated outright that costs should be incorporated into the model. Ten HTA organizations suggest using costs only for the conduct of a cost-effectiveness analysis.

Budget analysis done

Four HTA organizations provided guidance on conducting a budget analysis and all recommended a budget impact analysis as part of HTA.

Impact on project budget

AHRQ was the only HTA organization that addressed the impact of conducting a systematic review on a modeling project budget and suggested that “modeling efforts could easily consume 20 to 40 percent of the budget for a systematic review”.²

Conclusions

Mathes (2013),⁸⁴ upon which our search was based, summarized recommendations on methods for the preparation of economic evaluation by international HTA organizations. We

extended their review by summarizing the HTA organizations' recommendations within the framework laid out by Sculpher (2000),⁹ Philips (2004),^{6,216} and Kuntz (2013),^{2;217} as elaborated on by our systematic review of recommendations for decision and simulation modeling (*Chapter 1*). This structured framework, which allowed us to categorize recommendations into the 11 domains in **Table 2**, highlighted differences across HTA organizations with respect to the breadth and detail of guidance provided to modelers.

Despite the relatively large number of international HTA organization websites, only 17% (21 agencies) provided guidance regarding the application of decision modeling in the context of conducting an HTA. The majority of these HTA organizations are from Europe, Australia/New Zealand, and Canada. AHRQ provided the most guidance among the 21 websites followed by NICE and HIQA/Ireland. HTA organization mostly addressed how to model and how to integrate modeling into HTA.

The HTA organizations varied widely in what areas of modeling for which they provided guidance and what specific recommendations they made. This is consistent with the heterogeneity across researchers and organizations in the recommendations on the conduct and reporting of decision and simulation models, described in *Chapter 1*. For example, although 17 HTA organizations provided guidance on how to model, guidance did not consistently address the same themes (i.e., data, structure, consistency) regarding how to model. Moreover, while 15 HTA organizations commented on the integration of modeling as part of HTA, not all HTA organizations required a decision model as part of HTA development. Finally, variation in the frequency of reporting across the 11 subcategories may reflect the relative importance of these aspects of decision modeling to specific agencies and/or countries.

A number of practical constraints may have limited our review of HTA organization websites. We relied on a previous review of HTA organization websites,⁸⁴ which provided an apparently comprehensive listing of international HTA organization websites, but agencies not present on these lists may have provided additional modeling guidance. In addition, HTA organization guidance not provided on a website were not included. We considered using Google Site Search (<https://www.google.com/cse/sitesearch/create>) to more comprehensively screen individual websites to identify modeling guidance however, the often large number of potentially relevant items identified by this search tool made using this methodology infeasible. In addition, about one-third of the websites accessed (38 agencies) lacked an English language translation and could not be reviewed to identify additional relevant modeling guidance.

In summary, only a small number of HTA organizations (21 of 126) provide guidance for incorporating decision modeling alongside systematic reviews. Most HTA organization guidance provided recommendations about whether to incorporate models into HTA, generally favoring including models. Most also provided recommendations on how to model, focusing primarily on model data and structure, with fewer recommendations on model assumptions or validation. Similarly, most also recommended inclusion of costs in cost-effectiveness models. Few HTA organizations provided guidance related to other aspects of modeling in the context of HTA, including whether modeling should be conducted alongside systematic review, when modeling should be done related to the HTA, whether pre-existing models could be used, how systematic reviews should be incorporated into the modeling process, whether models should incorporate quality of life, whether a budget impact analysis should be done as part of the HTA, or who should conduct the modeling or the project budget impact of adding modeling to an HTA. Variability in recommendations probably reflects the heterogeneous needs addressed by HTA agencies operating in different jurisdictions. Harmonizing guidance across HTA agencies, and

adopting a common set of best practices whenever possible, would allow for more efficient transport of modeling results (or specific model implementations) across agencies.

Chapter 3. Future Research Needs For Decision and Simulation Modeling

Introduction

Trade-offs between benefits and harms are common in most clinical contexts and should be weighed against each other in decisionmaking. Providers, patients, and policymakers are increasingly interested in complementing evidence of benefits and harms with information from decision and simulation modeling to explicitly answer pressing policy needs. Modeling of healthcare conditions and management options, ideally based on evidence from systematic reviews, can provide a single, comprehensive, explicit and interpretable analysis of uncertainty,³ values and resource utilization to guide decisionmaking and to support the prioritization of future clinical research activities.^{4,5} Decision and simulation modeling theory offers a coherent approach for integrating clinical research evidence and patient values to optimize choices maximize expected utility for individual and for population health. However, gaps remain regarding how best to conduct and report decision and simulation models.

The systematic review and panel discussion described in Chapter 1 underscored the need for further research on ways to improve upon the performance of and uses for decision analyses in the context of systematic reviews. Here, we describe the stakeholders' prioritization of future research needs topics to advance decision and simulation modeling.

Methods

We conducted a systematic review of the published literature for evidence- and consensus-based guidance on the conduct of decision and simulation models. From the systematic review articles, we extracted suggestions for future research (both future research needs and research gaps) made by their authors. We then categorized the finding from the systematic review and convened an expert and stakeholder panel to discuss our findings and to prioritize future research needs in that context.

As described in Chapter 1, we formed a stakeholder panel that included experts in decision and simulation modeling, systematic review and evidence-based medicine, and epidemiology and biostatistics. The workshop also represented perspectives from six stakeholder groups¹²—patients and the public; providers of care; purchasers of care; payers, policy makers; and principal investigators, researchers and research funders—including policy makers, AHRQ Evidence-based Practice Centers, guideline developers, CISNET (Cancer Intervention and Surveillance Modeling Network), modelers, epidemiologists, statisticians, professional societies, payers, and patient advocates. **Appendix A** lists the 28 stakeholders who participated in the panel. The panel met in-person or remotely at a workshop at AHRQ on 27 February 2013. The goals of the workshop were to review and expand the list of recommendations and research needs identified by the systematic review of methodology recommendations for decision and simulation modeling, and to prioritize research to improve the usefulness and credibility of models to inform decisionmaking.

Stakeholders were encouraged to comment on individual recommendations, identifying gaps, limitations and areas for expansion. They then discussed future research needs topics. The

groups reviewed and discussed the list of future research needs gathered from our systematic review. The stakeholders were then asked to prioritize the list of future research needs derived from the systematic review. In addition, during the meeting, we solicited additional topics from participants, and following the meeting, the stakeholders prioritized these also.

To direct discussions about future research needs and deliberations about their prioritization, we asked stakeholders to consider four dimensions of need—Importance, Desirability of new research, Feasibility, and Potential impact—described more fully in **Table 4**. Due to methodological restrictions to comply with Federal policies, we used multiple methods with our stakeholders to assess their priorities. At the meeting, stakeholders were provided with a form listing the future research needs derived from our systematic review and were asked to rank order each on a scale of 1 (“Not desirable”) to 10 (“Essential”), with an option for no opinion. After the meeting, stakeholders were sent the list of additional future research needs proposed by the stakeholders themselves and were asked for their ratings; however, to prevent the request from being a survey, no specific method for rating was suggested. Most stakeholders who responded used a 1-10 scale but alternative methods used included categorizing into high, moderate, and low priority; 1 to 7 stars, and others. We therefore normalized all scales to a 10 point scale. Because of the different systems used to prioritize future research needs from the systematic review and the stakeholders, these two priority lists were kept separately. We selected the approximately five highest priority future research needs from each stage, using a natural breakpoint between higher and lower priority topics rather than a strict threshold of five topics.

Table 4. Dimensions of need (Approach to prioritization)*

Dimension	Definition
Importance	<ul style="list-style-type: none"> • Represents a critical uncertainty for decision makers • Advances credibility, transparency, and methodological rigor of modeling • Represents important variation or controversy in modeling practice • Represents high burden in time, effort, or resources to modelers
Desirability of New Research/Duplication	<ul style="list-style-type: none"> • Would not replicate ongoing or prior research, or established knowledge
Feasibility	<ul style="list-style-type: none"> • Ability to perform research
Potential Impact	<ul style="list-style-type: none"> • Potential for significant health or economic impact with clear implications for resolving important dilemmas in health and healthcare decisions or inequities or vulnerable population • Frequency (high frequency implies greater potential impact, and vice versa)

* Based on Standardized Selection Criteria of the Agency for Healthcare Research and Quality’s Effective Health Care Program.¹⁰⁸

Results

We reviewed 71 papers reporting recommendations for conducting decision and economic analyses, and simulation modeling. The future research needs and research gaps presented in these papers are summarized in **Table 5**. Suggestions for future research were made in 38 of the 71 papers (54%).^{6;11;15-17;19-22;24-27;29;30;32;36;37;39-41;43;44;48;50;52;53;56;58;59;61-63;65;66;69;71;74;216} Initial discussion among stakeholders focused on the purpose of prioritizing future research needs with reasons including using models to guide grant funding and the distribution of research and funding across diseases, intervention types, and methods, and choosing appropriate topics for

evidence review. Stakeholders felt that future research needs should span various model types. Aspects of model evaluation, including model validation and calibration, emerged as important targets for methodological research. Methodological advances in model evaluation were considered important for ensuring the validity and enhancing the credibility and acceptability of modeling results. Stakeholders presented multiple views across various future research needs. An example includes conflicting views expressed on the level of prioritization to place on using non-randomized trial data for assessing treatment effectiveness based on the existence of ongoing research on this topic.

One group of stakeholders discussed the use of multilevel modeling of primary or secondary (aggregate) data, with a particular focus on using data from multicenter/multiregional studies as an important field for future research. This group also noted that the topics identified by the systematic review as targets for future research were somewhat “Euro-centric”, possibly reflecting the origin of a large number of the publications included in the review. The panel suggested that effort should be directed toward identifying additional research priorities with an emphasis on the United States healthcare system and its needs. Stakeholders uniformly emphasized the importance of performing research on “widening the audience for using models,” including research on how to communicate the results of modeling studies to different audiences. Discussions concluded with the identification of 30 additional future research needs to be included in prioritization exercise (**Table 6**).

Based on stakeholder feedback, the 10 future research needs that were considered high priority by the stakeholder panel as a whole are highlighted in **Tables 5 and 6**. The highest priority future research need about model structure is a review of the standards for best practices in fields outside of medicine, such as engineering and environmental modeling. The goal would be to ensure that modeling of medical topics uses the most up-to-date practices and to determine how medical models could be improved by using, testing and adapting methodological advances from outside healthcare. This future research need was of particular interest to principal investigators, those most likely to develop decision models.

Another priority for future research pertained to incorporating surrogate outcomes in decision models, and to evaluating the assumptions invoked when using surrogate outcomes. Three future research needs were prioritized regarding where model data come from. Stakeholders wanted to better understand how nonrandomized trial data should be appropriately used, and if and how these data need to be assessed; how the bias inherent to many studies, regardless of design, ought to be handled; and what is the validity of, the indications for, and the best practices for conducting multiparameter synthesis.

Table 5. List of future research needs derived from the systematically reviewed articles

Future Research Needs	Prioritized Future Research Needs			
	All	Policy-makers	Principal Investigators	Providers
MODEL STRUCTURE				
Incorporation of surrogate outcomes, often done naively, and in a 1-to-1 relationship between surrogate and clinical outcomes	X		X	
Methods for assessing transferability/generalizability of economic analyses				
MODEL DATA				
Use of non-randomized trial data for assessing treatment effectiveness, including bias assessment, bias modeling/corrections, and selection modeling	X	X	X	X
Multiparameter evidence synthesis, particularly for parameters not related to treatment effectiveness (NB: this subsumes indirect/network meta-analysis)	X	X	X	X
Parameterizing models for probabilistic sensitivity analysis, particularly when data for a parameter are sparse				
Develop and standardize techniques and processes for structuring complex models in the setting of HTA that are accessible to decision-makers				
Formalizing methods for interpreting non-probabilistic sensitivity analysis				
Conceptualizing the search process for parameters other than treatment effectiveness; developing practical methods for searching, including standardized				
Use of multi-level modeling of primary or secondary (aggregate) data, with a particular focus on using data from multi-center/multi-region studies				
Assessing willingness to pay and developing conversion factors (for health outcomes)				
MODEL CONSISTENCY				
Determine optimal methods for model validation and calibration	X	X	X	X
Assessment of structural uncertainty	X		X	
<i>Error research (research on methods for preventing and identifying errors in the modeling process)</i>		X		
Methods for automated model checking (for structural errors as well as logical/numerical errors)				
Methods for assessing indirect costs (e.g. for individuals outside the labor force)				
RESULTS REPORTING AND INTERPRETATION				
Implementation research on the application, use, and impact of modeling (are decisions/outcomes improved)	X	X	X	X
Work on developing flexible and comprehensive systems for evaluating completed economic analyses				
<i>Relationship between uncertainty and inference (should decision-makers “trade” level of uncertainty against cost-effectiveness)?</i>				X
Determining threshold values for the incremental cost-effectiveness ratio				

Future research needs prioritized by all stakeholders are in bold; those prioritized by only policymakers, principle investigators, or providers are in italics. Within categories, future research needs are listed in the order of prioritization by all stakeholders.

Table 6. List of future research needs derived from stakeholders

Future Research Needs	Prioritized Future Research Needs			
	All	Policymakers	Principal Investigators	Providers
MODEL STRUCTURE				
Review of standards for best practice in fields outside medicine (e.g., engineering, operations research, environmental modeling, etc.)	X		X	
<i>Research on using decision models in decision aids for shared decision-making</i>				X
<i>Review of standards for best practices in the development of decision analysis and simulation models for patient-centered comparative effectiveness questions</i>		X		
Research on the use of duplicate modeling (building independent models with common inputs) to explore structural uncertainty				
<i>Research on individualizing models – predictive value for individual patients</i>				X
How to build decision models to illustrate trade-offs in patient-centered outcomes				
<i>Developing multi-purpose/multi-disease models; research on “reusable models”, models that can be repurposed to be used for different decision problems that fit under the same model structure</i>		X		
Identifying cases where relatively simple models are “good enough” for guiding decisions – for example, exploring cases where simple and more elaborate models agree or disagree, to identify patterns				
<i>Research on the choice of “appropriate” modeling approaches for different decisionmakers (considering policy vs. patient-level decisions and the trade-off between complexity and transparency)</i>				X
Review of standard for best practice within more specific topics (e.g., specific model types)				
<i>Computational complexity for some modeling is too high – and there is computer science or applied math approaches that could be explored</i>			X	
Research on multi-level modeling for populating models, with particular focus on using multi-level models to reflect variability in patterns of care				
Framework for deducing the sufficient complexity of a model for a given question				
When to model, what to model, how to model				
Framework for pushing the use of conceptual modeling as the first step -- to help understand what’s important and what is not				
MODEL DATA				
Modelers often “take data as they are” and plug them in; however we can break down the variability in the data into sampling error, bias, and heterogeneity; we need better understanding of the role of bias, and of how to handle it	X		X	X

<i>Emphasis on “not is it cost-effective” – but for whom is it cost-effective? In most cases this will involve the use of individual participant data</i>			X	
How to account for distributional justice/equity/utility tradeoffs (benefits for some are not accrued by others)				
Research on appropriate measures of economic value (i.e., without focusing exclusively on willingness-to-pay)				
What is a structural sensitivity analysis in one modeling approach is a parameter in another – understanding this duality is important as it is easier to handle the latter				
Methods for multidimensional utility assessment (e.g., a joint utility for treatment and outcome sequences, a joint utility for combinations of morbidities versus combining separate utilities for single morbidities)				
MODEL CONSISTENCY				
Methods for accounting for heterogeneity including baseline risk and benefit, health status, and individual patient preferences	X	X		
<i>Development of methods to use simulation models to address questions on heterogeneity of treatment effect</i>		X		
Assess quality and applicability of models				
Performing cross model comparison and selecting a model				
RESULTS REPORTING AND INTERPRETATION				
Optimal methods of communication of models to end-users; additional education needs for communication; how models are used and communicated; Widening audience for using models – research on how to communicate results to different audiences	X	X	X	X
OTHER				
<i>Methods to engage and tailor methods and objectives to end users</i>				X
Methods for using value of information to choose among a broad range of alternative study designs for different interventions				

Future research needs prioritized by all stakeholders are in bold; those prioritized by only policymakers, principle investigators, or providers are in italics. Within categories, future research needs are listed in the order of prioritization by all stakeholders.

Stakeholders prioritized three future research needs regarding model consistency. The highest priority among these is to determine which methods for model validation and calibration^a are most appropriate, most improve the validity and applicability of models, and which methods are most likely to be feasible for use. In addition, which methods should be used to examine the impact of alternative model structures and when this should be done.

Stakeholders also prioritized research into methods for accounting for heterogeneity within models, including baseline risk and benefit (or treatment heterogeneity), health status, and individual patient preferences. The stakeholders' logic, particularly policymakers, was that, in general, the most useful models are those that can be individualized for particular patients.

Implementation research on the application, use, and impact of modeling was identified as a future research priority for results reporting and. The goal of this research would be to determine how models can be framed and presented to maximize their value for real-world decisionmaking. Stakeholders also prioritized research into optimal methods to communicate models to end-users, including education of end-users, how models are presented, and how to fulfill the needs of different audiences.

It is worth noting that policymakers and providers, in particular, prioritized several different future research needs than the stakeholder panel as a whole, in line with their particular perspectives. Policymakers prioritized future research needs that address ensuring that models are accurate, patient-centered, re-usable, and address heterogeneity. Specifically, these included research on methods for preventing and identifying errors in the modeling process, best practices for developing models for patient-centered comparative effectiveness questions, developing multi-purpose and multi-disease models that can be repurposed, and methods to use simulation models to address questions on heterogeneity of treatment effects.

Providers' highest priority future research needs generally revolved around how to individualize models for patients and how to make them most useful in clinical practice. Specifically, these included research into the relationship between uncertainty and inference—how decision-makers should trade-off level of uncertainty against cost-effectiveness, how to use decision models in decision aids that are used for shared decision-making, how to individualize models and to provide predictive value for individual patients, and how to engage and tailor the model methods and objectives to end-users.

Principal investigators, in contrast, tended to prioritize future research needs regarding the mechanics of developing models. In addition, to the future research needs described above on assessment of structural uncertainty and best practice standards from outside of medicine, principal investigators also prioritized research into how to handle models for which the computational complexity is too high to develop and how to incorporate individual participant data into models to determine for which people are interventions cost-effective.

^a Validation and calibration are methods to test how the models comport with reality as measured by empirical data. As such they can provide information that enhances model credibility and acceptability as well as provide insights into the potential use of modeling decisions in practical settings.¹⁰⁹

Conclusions

Based on a systematic review of 71 publications providing recommendations for the conduct and reporting of decision and simulation models, we summarized a list of future research needs and presented these statements to a diverse stakeholder panel with expertise in decision and simulation modeling, systematic review and evidence-based medicine, and epidemiology and biostatistics. In addition, we solicited additional future research needs proposed by these stakeholders, who prioritized both lists of future research needs. The stakeholders prioritized future research needs with the goals of improving the methodology for the conduct of modeling, the validity of the models, and the communication of their findings. The future research needs were also prioritized primarily from the stakeholder perspectives it was believed most commonly directly use healthcare models, namely, policymakers, principal investigators and physicians. Other stakeholders (e.g., patients) were considered, but were thought to have a lesser direct impact from research into improving modeling methodology. The prioritized future research needs principally involved questions about model data, model structure, consistency, and reporting.

First, perhaps in recognition of the imminent “big data” revolution in healthcare that can be characterized by the volume, complexity, diversity and timeliness of data, our stakeholders acknowledged the need to assess the appropriate use of non-randomized trial data for determining treatment effectiveness, including the potential for bias assessment, bias modeling/corrections, and selection modeling.¹¹⁰ For example, the Food and Drug Administration’s Sentinel Initiative has led to its Mini-Sentinel Initiative and Observational Medical Outcomes Partnership which have now blossomed into the Reagan-Udall Foundation (RUF),¹¹¹ an independent nonprofit public-private collaboration to generate post-marketing evidence from huge heterogeneous data sources of the use of FDA-regulated drugs, devices and procedures in the real-world so that the healthcare community can reliably identify harms and opportunities to improve patient care. Moreover, across health technology assessments agencies,^{84;112} recent reviews concerning the use of observational versus randomized trial data have identified conflicting recommendations with one agency preferring observational data.

Second, much work is needed to better understand how data from different sources—including randomized trials, other trials, database analyses, observational studies, epidemiological data—should be used and assessed and, possibly, adjusted for risk of bias. Current guidance on how to handle data from multiple sources relies on transparency, researcher judgment, and assessment of uncertainty.^{6;216} However, modelers would benefit from better evidence on how and whether to use, assess, and adjust for potentially biased data from multiple sources. The other related future research need considered multiparameter evidence synthesis, particularly for parameters not related to treatment effectiveness. Of note, the issue of incorporating multiparameter evidence subsumes indirect/network meta-analysis, an analysis tool for which there is a need for guidance about its use in models. Using Bayesian methods and Markov Chain Monte Carlo software, this approach synthesizes a broad range of alternative evidence sources, but can also examine the consistency of the evidence provided by these multiple information sources.¹¹³ Stakeholders discussed the need for “chains of evidence” reasoning to piece together disparate pieces of evidence, such as evidence on intermediate and terminal outcomes, or evidence from different study designs. Having quantified the uncertainty in the underlying evidence base, such types of analyses can also be used to prioritize future research by examining the impact of reducing uncertainty. Although examples of multi-

parameter evidence synthesis exist, interest has grown as demonstrated by a seven-part tutorial in *Medical Decision Making*.¹¹⁴⁻¹²⁰

Third, regarding model consistency—specifically validation and verification methods—to improve model acceptability, models need to be validated and calibrated to ensure the credibility of modeling results.^{78;121} Multiple methods are available for both validation and calibration, but there is limited evidence comparing specific methods.⁵ Although CISNET colon cancer models have been systematically compared,¹⁰⁷ studies are needed specifically to address which validation or calibration methods and approaches are most appropriate for alternative types of models for different diseases. Calibration methods vary in their time and resource requirements; thus, the most appropriate method may not be the “best” method in all circumstances. Lastly, increasing scientific journals have called for reproducible research as a foundation for scientific evidence.^{78;122;123} In modeling this would consist of cross model (between model) validation where independently produced models yield similar results (convergent validity).¹²⁴

High profile journals such as *Science*,¹²⁵ have called for shining light into computational science, i.e., black boxes. As articulated by Weinstein, “Decision makers will not readily accept results and cost-effectiveness unless they can understand them intuitively and explain them to others in relatively simple terms.”¹²⁴ Consistent with these trends, the stakeholder identified optimal methods of communication of models to end-users; additional education needs for communication; how models are used and communicated; widening audience for using models – research on how to communicate results to different audiences as a future research need.

Lastly, in an upcoming era of value-based payment for outcome, stakeholder prioritized the need to explore implementation research on the application, use, and impact of modeling (are decisions/outcomes improved). Models are commonly accepted in decision-making in such fields as environmental protection, weather prediction, and defense strategy, but less so in healthcare.¹²⁶ Better use of up-to-date methods used in these and other fields could only improve medical models. One challenge is such assessments, “In our view, the most important thing to keep in mind in evaluating a health-care evaluation model is that its outputs must not be regarded as claims about the facts or as predictions about the future. Rather, its purpose is to synthesize evidence and assumptions in a way that allows end users to gain insight into the implications of those inputs for valued consequences and costs.”⁷⁷

In conclusion, this systematic review and expert panel provides a comprehensive collation of methodological guidance developed through various methodological processes and identifies ways to improve upon and standardize the use of decision analyses in the context of systematic reviews. This review updates previous syntheses of evidence- and consensus-based guidance on the conduct of decision and simulation models and the stakeholder panel prioritizes future research topics needed to advance the current state-of-the-art in decision and simulation modeling.

Chapter 4. A review of validation and calibration methods for healthcare decision and simulation models

Introduction

In practice, models of at least moderate complexity will be ‘solved’ with computer-based numerical analysis and simulation. Because these computer models are used to inform predictions or decisions in the real world, assessing their credibility (trustworthiness) is paramount. A recent ISPOR-SMDM Good Research Practices Task Force identified model validation as one of the two determinants of confidence in models (the other being transparency).⁷⁸ Further, because some aspects of reality are unmeasured or unknowable, models will often require inputs for which no or only partial data exist. In such cases, model calibration can be used to select input values that lead to model outputs “as close as possible” to available empirical data. Because model validation and calibration entail a “confrontation of models with data” they can inform judgments about the credibility of models, and can guide the use of modeling results in practical settings.

Our systematic review of evidence- and consensus-based guidance on the conduct and reporting of decision and simulation models (*Chapter 1*) identified model calibration and validation as a major methodological research area. This was triangulated by a panel of multiple stakeholders, including developers and users of healthcare decision models, who also identified aspects of model evaluation, and model validation and calibration in particular, as important targets for future methodological research. There is limited previous work surveying calibration and validation methods and most existing reviews have either focused on a limited topic area (e.g., treatment of cardiovascular disease, cancer natural history) or modeling methodology (e.g., micro-simulation models).

Based on the above, we conducted a project aiming to provide a unifying overview of validation and calibration methods, and a survey of studies comparing validation and calibration methods used in healthcare decision and simulation modeling.

Methods

Sources of information and review methods

Issues pertaining to model evaluation and assessment arise in many methodological areas (e.g., mathematics and statistics, economics, theory of simulation, operations research, management science), as well as many topic areas (e.g., healthcare, disease modeling, biology, environmental science, mechanical engineering, material science). The relevant literature is vast, poorly categorized in standard literature databases (i.e., specific search terms are lacking), and published as journal papers, conference proceedings, books, and technical reports that are not always easy to identify, obtain, and comprehend. Thus, a comprehensive systematic review of all relevant methodological papers was deemed non-feasible. Instead, we relied on a mixed approach that combined consultation with expert methodologists; hand-searching the reference lists of related papers, technical reports, and books; review of our personal reference collections;

and a systematic review of studies comparing validation and calibration methods for disease modeling-related or healthcare-related models.

The systematic component of our literature review covered four electronic databases (MEDLINE, Cochrane Methodology Register, Health Technology Assessment Database, NHS Economic Evaluation Database), through June 3, 2013, for articles presenting validation and calibration methods in reports of decision or simulation models. We also rescreened the citations retrieved by the search strategy of our recently completed systematic review of recommendations for the conduct and reporting of decision and simulation models. The final search strategy, with the list of 37 included journals, is presented in **Appendix B**.

Six reviewers independently screened 6825 abstracts in duplicate and resolved disagreements by group consensus. Eligible studies had to compare or apply at least two methods related to model validation, model calibration, or goodness of fit, in the context of a decision model. We excluded studies that applied only a single method of validation or calibration. Three reviewers extracted descriptive information from included articles. Completed data extraction forms were verified by a second reviewer and were discussed during group meetings. Extracted data included population characteristics, outcomes, basic model description, and methods and results related to model validation and calibration.

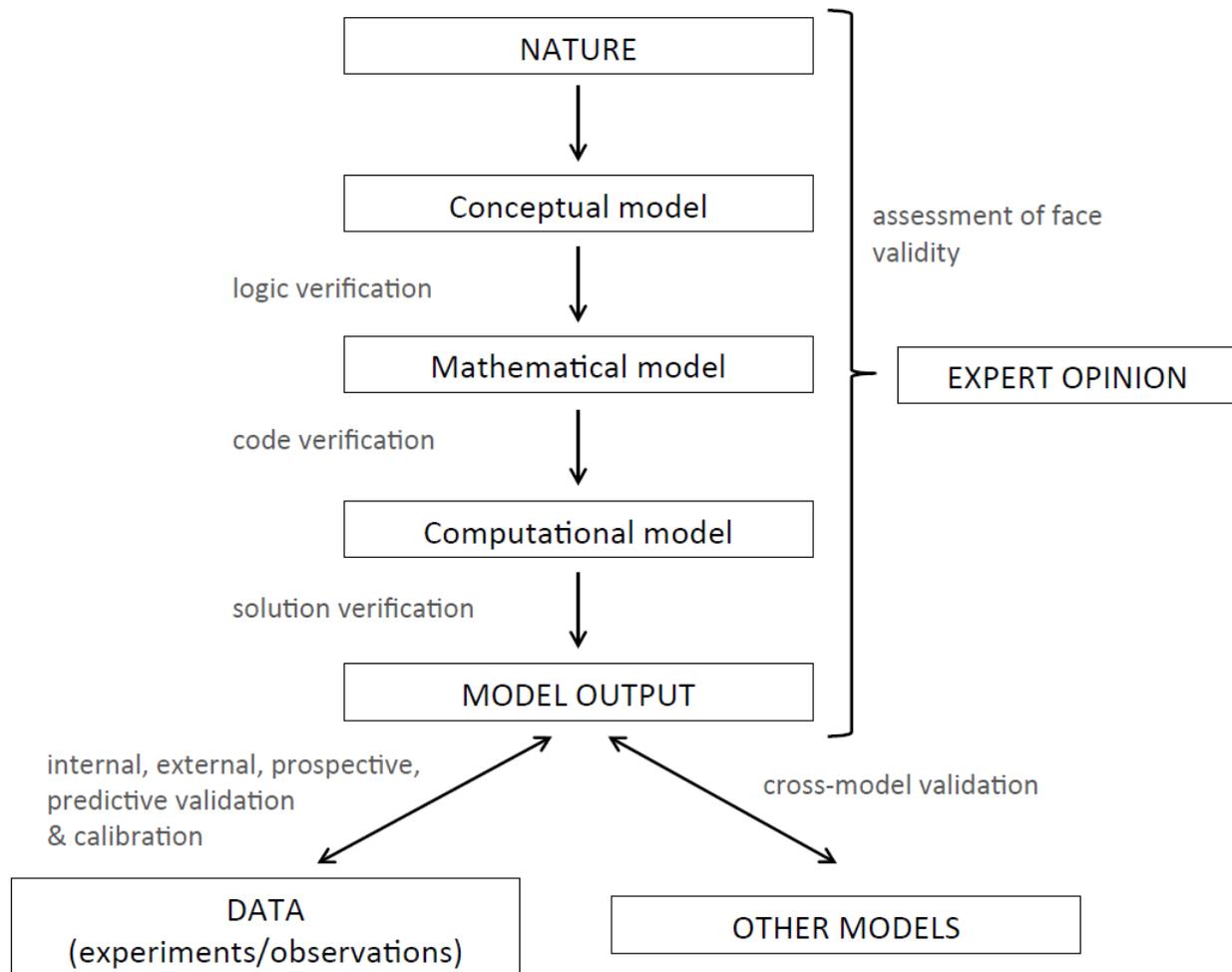
Definitions and preliminaries

By ‘modeling’ we mean a multistep iterative process to conceptualize an abstraction of salient aspects of reality (conceptual model), specify it mathematically (mathematical model) and implement it in computer code (computational model) so that it can be ‘solved’. **Figure 3** outlines the modeling process. Because the modeled natural phenomena can be complex and the model implementation process is often intricate, it is important to perform checks throughout. Terminology about these checks varies across fields and topic areas, but the underlying concepts are similar.

In this work we use the terms ‘assessment of face-validity’, ‘verification’, ‘validation’ and ‘calibration’ to describe various checks. **Face-validity** refers to whether the model is deemed a satisfactory representation of the salient aspects of reality and whether the model results appear to be plausible. **Verification** refers to assessing the correctness of the mathematical structure (e.g., absence of mistakes in the logic), and of the implementation of the computational model (e.g., absence of software ‘bugs’, suitability of numerical algorithms). We use the term **calibration** only for the process of determining the distribution of unobserved (possibly unobservable) parameters so that model outputs match (i.e., “fit”) observed empirical data. We define **validation** to be the comparison of model outputs with expert judgment, observed data, or other models, without any attempt to modify model parameters to improve fit.

Figure 3. Diagram of the modeling process

Model evaluation & assessment



Results: Overview of Validation and Calibration Methods

Model validation

Validation is the assessment of the “congruence” between model predictions and actual observed data, or the results of other models addressing the same (or similar) research question, or expert predictions of what the results should be. The literature identifies several aspects of model validation, including face validity, verification (internal validation), cross-model comparisons, external validation, and prospective and predictive validation.^{50;78;127;128} To a large extent, the definitions provided below follow those adopted by the recent ISPOR-SMDM Modeling Good Research Practices Task Force.⁷⁸

- *Face validity* (“first order” validation) refers to the determination, by a suitable group of experts, that the model reflects the current understanding of the science and available evidence. Expert review should cover all aspects of the modeling process, including the

question formulation, model structure, model data, and the model output. Evaluation of aspects of modeling other than the model output, are best performed “blinded” to the model results to reduce the possibility of biased assessment.

- *Verification and internal validation* (“second order” validation) refers to an assessment of whether the model has been implemented correctly and behaves as expected. Verification can pertain to the computer code (“code verification”) and the solutions that it produces in well-understood problems with known solutions (“solution verification”).¹²⁷ Some consider the comparison of model outputs with the data used to populate the model to be a component of internal validity. However, when using a study to estimate model parameters only a small part of the study results are used; in which case other data can be used for validation. For example, if a randomized trial is used to inform a model regarding the relationship between treatment (e.g., statin vs. no statin) and a surrogate outcome (e.g., LDL cholesterol at 6 months), the trial data on other outcomes (e.g., mortality, morbidity) can be used for model validation.⁷⁸
- *External validation* (“third order” validation): In external validation, the model outputs are compared to empirical observations that were not used in model development. As noted above, external validation is sometimes taken to mean the comparison of model outputs against observations in datasets that are disjoint from those used in model development. In general, external validation using disjoint data provides a more stringent test of model performance.
- *Prospective and predictive validation* (“fourth order” validation) assess the model’s ability to reproduce (“predict”) empirical results that were not available and were not used during its development.¹²⁸ Prospective validation refers to the use of data that accrues over additional followup in studies that were used for model development. Predictive validation refers to the use of data from independent studies that were unavailable at the time of model development.
- *Cross-model validation* involves a comparison of results among different models for the same (or sufficiently similar) analyses.⁷⁸ Such comparisons can increase the credibility of the models and provide methodological insights. Cross-model comparisons have been used extensively in cancer simulation models supported by the Cancer Intervention and Surveillance Modeling Network (CISNET).¹²⁹

Two general types of methods are in common use for internal, external, and prospective or predictive validation: ‘informal’ methods using graphical and tabular presentations of model results (e.g., time series and scatter plots, cumulative frequency distributions) and ‘formal’ (statistical) using a distance function or goodness-of fit metric.¹³⁰ The value of graphical and tabular data displays cannot be overemphasized. However, these methods may not always have adequate sensitivity for detecting poor fit to the data and are operator dependent. For this reason, graphical methods are usually combined with statistical methods. The latter rely on assessments of goodness-of-fit that quantify the discrepancy between observed data and model outputs,^{131;132} in many ways these quantitative assessments are similar to standard model fit criteria used in statistics.¹³³ When a Bayesian approach is adopted, posterior predictive checks (i.e., comparisons between the observed data and the model’s posterior predictive distribution) can be used to assess model fit.^{134;135} For example, Gardner (2011) proposed and studied alternative model fit statistics for individual-level infectious disease models, based on posterior predictive checks.¹³⁶

It is not possible to identify a universally preferred method for statistical model validation.¹²⁷ However, some general principles to guide the choice are identified in the literature: first,

statistical validation should be applied to quantities of interest that are relevant to the model scope and the perspective of the analysis. Second, statistical criteria for model fit need to be appropriate for the mathematical structure of the model.¹²⁷ For example, if there is dependence (clustering) in the statistical model (e.g., repeat measurements within individuals or groups) the statistical criteria should account for that dependence. Third, because model fit is generally improved by increasing the number of model parameters (which may lead to over-fitting the data used for model development and limit generalizability), criteria that “account for” the number of model parameters may be preferable. Fourth, any statistical method for model validation should take into account uncertainty in both the empirical data and the model outputs.

More generally, it is not possible to establish model validity in the affirmative; i.e., there is no criterion that, if met, establishes a model as generally valid. In fact, some experts suggest that it is only possible to demonstrate model invalidity in a specific setting and for a specific purpose (e.g., by showing that model predictions do not fit a set of observations that is relevant to the anticipated model use).⁷⁸

Our searches for studies using alternative model assessment methods identified several studies that applied various approaches to assess model fit.^{128;136-143} This list is obviously not exhaustive, in that it is limited to healthcare applications and is based on a systematic, but non-comprehensive process for identifying relevant studies. We found that validation methods were specific to the research question that was addressed by each study and the investigators’ choice of methods for model implementation. For these reasons, we did not use them to draw general conclusions. Nonetheless, these specific healthcare applications provide insight into evolving methodological research and standards.

Model calibration

Calibration involves the optimization of a subset of model parameters to improve the fit of model predictions to empirical data. Traditionally, calibration is distinguished from other estimation tasks by its use to obtain estimates for parameters that are inherently unobservable or for which no data are available (i.e., parameters that cannot be identified from the available data). For example, in microsimulation models of cancer in humans, unidentifiable parameters (e.g., growth rates of preclinical cancers) are determined by selecting model input rates so that model outputs (which are functionally dependent on the unidentifiable parameters) “are as close as possible” to empirically observable data.

Calibration efforts are tailored to the specific needs of a particular decision or simulation model. Calibration is fundamentally an optimization (estimation) problem. To specify a *calibration problem*, one has to define the following components:

- *Calibration parameters* are the typically unidentifiable or weakly identifiable (e.g., supported by imprecise evidence) model parameters that are subjected to calibration.⁴³ A most important issue is whether the *feasible domain of the calibration parameters is convex* or not. This is because convex calibration problems (problems where the parameter domain and the objective functions are convex) can be solved easier than non-convex ones.
- *Calibration targets* are the data against which the model output is compared; calibration aims to select parameter values that produce model outputs that are “close” to the calibration targets (while “close” may be assessed graphically or visually, it is preferably encoded quantitatively by an objective function). The choice of calibration targets

depends on the model quantities of interest, the availability of “high quality” data, and the goals of modeling. For example, when calibrating a decision model, the calibration targets should be data relevant to the decisional context, and obtained from well-designed and conducted studies of populations “similar” to those who will be affected by the decision.

- *Objective functions* are typically scalar functions of the calibration parameters used to assess “closeness” quantitatively. Typical choices include a *distance* of the calibration target data from model outputs, a convexity-preserving transformation of a distance, or a *likelihood* or *pseudo-likelihood*.^{131;132;144-146} Example distances are the sum of absolute or squared differences between model outputs and calibration targets (L1- and L2-norms, respectively). Examples of convexity-preserving transformations of distances are various chi-squared statistics. Distances are convex objective functions, and for many problems the likelihood and pseudo-likelihood is convex as well. As mentioned above, convexity of the objective function is an important property, because convex problems are much easier to solve than non-convex ones.

The goal is to solve (optimize) the calibration problem, that is, identify the feasible values of the calibration parameters that optimize the objective function. Solving the calibration problem entails defining the following:

- *Algorithm for optimizing* the objective function. These algorithms search for values of calibration parameters in the feasible domain that optimize the objective function. Principled searching uses mathematical programming to obtain values. Descriptions of *ad hoc* approaches, such as ‘manual’-tuning, however, also occur in the healthcare literature.
- *Acceptance criteria* are used to determine whether an algorithm has converged to a solution. Typically, this means that further iterations do not change the value of the objective function and the estimated values of the calibration parameters beyond prespecified tolerances (strict tolerances can be of the order of machine precision).
- The *stopping rule* is the criterion for terminating the calibration process. Usually, calibration is stopped when the acceptance criteria are satisfied, the search space is exhausted (e.g., all points in a grid search have been evaluated), or a predetermined maximum number of iterations has been reached.

As mentioned already, the solution of the calibration problem (identifying the global optimum of the objective function within the feasible domain of the calibration parameters) depends greatly on the objective function and the shape of the feasible domain of the calibration targets. When the problem is convex (both the objective function and the feasible domain are convex) or can be restated to be convex, a single optimum exists and the mathematical programming methods to find it are very robust, representing a readily-usable technology.¹⁴⁷ Problems that are not convex—or cannot be recast as convex—have local optima, and demand global optimization approaches. For such problems, exact solutions often become computationally expensive, and only approximate solutions may be practical.

Based on the above, we make the following general observations, which are important for interpreting the empirical research found in the healthcare literature:

- Judging which specification of the calibration problem (in particular, which objective function) is most appropriate is not answerable from data alone. This is a general

statement for optimization problems. The choice of the objective function should reflect the decisionmakers' perspective, and the nature of the problem.¹⁴⁸ Thus comparisons of solutions to alternative objective functions are difficult to interpret.

- From a theoretical basis, solutions to different specifications of the calibration problem clearly need not be identical. We interpret empirical demonstrations of this phenomenon in specific applications as stability analyses.
- Once a calibration problem has been specified, it is straightforward to rank the different optimization algorithms according to their performance, by ordering (within machine precision) the value that the (scalar) objective function has with each algorithm's solution. To learn from such comparisons one must (a) be confident that the algorithms were implemented correctly and efficiently; and (b) be able to characterize salient mathematical attributes of the calibration problem, e.g., whether common regularity conditions were met.^{147;149;150} Because this requires a very deep understanding of the problem at hand and of the mechanics of the various algorithms, it is generally not possible to draw generalizable conclusions.

Our search for studies comparing alternative calibration methods for healthcare models identified four relevant studies (**Table 7**). They include comparisons between different specifications of the calibration problem solved with the same algorithm, or with different algorithms; and comparisons between alternative algorithms for the same calibration problem, that highlighted various mathematical aspects. Further, each empirical study was limited to a single modeled process (and the same model was examined in 2 of the 4 studies). Many reported results (e.g., running time, performance, etc.) are expected to be dependent on the specific computer implementation. Thus, we deem that it is not possible to draw general conclusions from these studies.

Methodological appraisals of validation and calibration methods in health-care models

Several systematic methodological appraisals of healthcare-related decision and simulation models provide information on the validation and calibration methods used in practice. These studies have found that more than half of all modeling studies do not report any use of calibration or validation methods.^{43;151-162} When some aspect of model validation or calibration is mentioned, reporting is often incomplete. Below, we review the results of methodological appraisals that provided a more in depth assessment of validation and calibration methods.

Cancer

Stout (2009) reviewed 131 studies of cancer microsimulation models that could have used calibration methods to determine input values for unobservable parameters.¹⁵¹ Approximately 50% of the studies (n=66) referred to "calibration" or "model fitting" and an additional 16% (n=21) provided references to methodological publications on model calibration. Nearly all studies (95%, n=83 of 87) identified the calibration targets they used, 54% (n=47) reported information on the goodness-of-fit metric used. Information on the search algorithms used was not well described. The authors used the results of this investigation to derive a 7-item "*Calibration Reporting Checklist*".

Table 7. Studies comparing alternative calibration methods applied to the same problem

Author, year	Area of application	Model structure	Methods compared	Study findings
Kong, 2009 ^{163,164}	Lung cancer development, progression, detection, treatment, and survival (Lung Cancer Policy Model)	Agent-based; state transition model; 1 month cycle length	<ul style="list-style-type: none"> Search algorithms (simulated annealing vs. genetic algorithm) 	<ul style="list-style-type: none"> Both search algorithms attained study-determined threshold GOF scores within 1000 search iterations SA outperformed GA The model predictions after calibrations matched other mathematical models of cancer development
Taylor, 2010 ^{165,166}	Cervical cancer epidemiology, natural history, and effectiveness of vaccination	Cohort-based; 6-state Markov model; 6-month cycle length; lifetime horizon; implemented in Excel with Visual Basic for Applications.	<ul style="list-style-type: none"> Search algorithms ('manual' calibration vs. random search of parameter domain vs. Nelder-Mead) 	<ul style="list-style-type: none"> The Nelder-Mead algorithm and manual calibration achieved the best fit (weighted mean percent deviations of 7% and 10%, respectively); random search performed poorly (weighted mean percent deviation of 39%) Use of the Nelder-Mead algorithm required less analyst time but was more computationally demanding, compared to manual calibration.
Kamon, 2011 ^{167,168}	Choice of adjuvant chemotherapy for early breast cancer	Cohort-based; Markov model; 1 year cycle length; 50-year time horizon; implemented in Excel with an add-on component (Microsoft Excel Solver; Frontline Systems).	<ul style="list-style-type: none"> GOF metrics (chi-square vs. likelihood) Search algorithms (random vs. gradient-based guided search) Alternative convergence criteria (narrow vs. broad) 	<ul style="list-style-type: none"> The chi-square GOF metric "differentiated between the accuracy of different parameter sets" to a greater than the log-likelihood statistics The guided search strategy produced results of higher accuracy and greater precision than random search Broader convergence criteria produced less accurate results that were closer to the non-calibrated results
Taylor, 2012 ^{165,169}	Cervical cancer epidemiology, natural history, and effectiveness of vaccination	Cohort-based; 6-state Markov model; 6-month cycle length; lifetime horizon	<ul style="list-style-type: none"> Alternative starting values for the Nelder-Mead search algorithm (5, randomly chosen) GOF metrics [weighted MPD with weights for the cancer incidence and mortality parameters that were 6- and 3-fold larger than those of corresponding carcinoma in situ endpoints (1-6-3 weights) vs. MPD with 1-3-3 weights vs. MSPD with 1-3-3 weights vs. MSPD with 1-3-6 weights vs. ML] 	<ul style="list-style-type: none"> The sensitivity/stability analyses to the choice of initial values and alternative weighting schemes revealed a substantial amount of uncertainty in the model output – far greater than that revealed by forward propagation of uncertainty

GOF = goodness-of-fit; ML = maximum likelihood; MPD = mean percentage deviation; MSPD = mean squared percentage deviation.

Cardiovascular disease

Haji Ali Afzali (2013) reviewed 81 model-based studies (including cohort and agent-based models) for cardiovascular disease.¹⁵⁸ They found that 73% (59 studies) reported some element of model evaluation, but only 6% (5 studies) reported a calibration process. Usually multiple calibration targets were employed in each study but only a single study provided information on the goodness-of-fit metric and no studies reported information on the acceptance criteria. Search algorithms were generally not well documented.

Unal (2006) reviewed the methodology of 42 coronary heart disease models (reported in 75 publications).¹⁶¹ In general validation and calibration methods were not used systematically and were not reported in detail. Six of the 42 models were considered “principal coronary heart disease models” – of these, two reported some calibration procedure and only one reported the performance of model validation.

Neurological disease

Siebert (2004) reviewed 8 studies using mathematical models to evaluate treatments for Parkinson’s disease.¹⁶² None of the eight studies reported any internal or external validation of their models. Dams (2011) surveyed 11 cost-effectiveness studies for Parkinson’s disease including therapeutic and diagnostic evaluations.¹⁵⁹ They found that only four models reported performing some form of model validation and none provided adequate details of their validation methods and results.

Respiratory disease & smoking cessation

Ferdinands (2008) reviewed 13 disease simulation models of asthma or chronic obstructive pulmonary disease (11 state-transition models and 2 dynamic population models).¹⁶⁰ Only two studies provided information on code and solution verification; seven studies reported comparisons of model outputs with data used to develop the model; seven studies reported results of external validation; and no studies reported performing predictive validation or plans to undertake such efforts.

Bolin (2012) assessed 78 economic evaluations of smoking cessation therapies,¹⁵⁵ 30 of which were considered “highly relevant” (defined as studies applying “intertemporal modeling with a time horizon” of at least 20 years). They found that “several studies”^b used simulation models – that were not described as previously validated – without performing any model validation.

Calibration as estimation

As described, the calibration process is very similar to statistical estimation. Both processes have the same goal, namely to find input values that lead to the best possible model fit. For example, if the objective function of the calibration is a likelihood function, calibration is—by most any definition—a statistical estimation procedure. We explain that this conceptualization is important for assessing the consistency of data inputs, and for recognizing the extent of nonidentifiability of the parameters of the mathematical model.

^b An exact count was not provided in the main text of the paper and the supplementary appendix was not downloadable from the journal Website.

The description is easier for simulation models that use meta-analysis to inform some of their inputs. First, note that the empirical data inputs for the model comprise two potential approaches and two types of data (1) *meta-analysis-estimation* of input data to estimate some model parameters, and (2) *calibration-estimation* in which calibration targets are used to estimate the remaining model parameters. Modelers have two options: do the meta-analysis-estimation and the calibration-estimation separately (as two steps; most common practice) or jointly (in one step; least common). We make the following observations about one-step versus two-step procedures:

- Compared to one-step estimation, *two-step estimation is generally more inefficient (in the statistical sense) and does not guarantee that the best-fitting values for parameters will be identified*. It can also hinder the complete characterization of parameter uncertainty and the representation of correlations between data sources or dependencies among model parameters. The one-step method is consistent with the scientific maxim of using all available evidence when making decisions. Further, it may help avoid under-assessments of uncertainty. The one-step approach is closely related to methods for synthesizing evidence from diverse sources, including multi-parameter and generalized evidence synthesis,^{113;170} the confidence profile method,¹⁷¹⁻¹⁷⁴ cross-design synthesis,¹⁷⁵⁻¹⁷⁷ and teleo-analysis.¹⁷⁸
- *One-step estimation allows for formal tests of consistency of parameter estimates obtained by different sources of evidence*. One-step approaches enable an assessment of whether the various data sources ‘square up’. If the data are inconsistent (do not ‘square up’), a serious problem exists that requires resolution (discussion of possible methods for resolving inconsistencies is beyond the scope of the current work).^{83;179} If the data are consistent, the one-step approach maximizes use of all of the available information.
- *One-step estimation allows one to use well-established quantitative methods for comparing differences between model outputs and empirical data while using all available data*.¹⁸⁰ Examples of such methods include posterior predictive checks, posterior mean deviance statistics, and various model cross-validation approaches.
- Under some circumstances, which can be formalized, the one-step approach and the two-step approach (as described above) are mathematically equivalent.^c

Parameter identifiability

The ability of Bayesian methods to incorporate external information or subjective beliefs, in the form of informative prior distributions, is particularly appealing when some model parameters are unidentifiable. For example, Rutter (2009) used a Bayesian approach to calibrate a microsimulation model of colorectal cancer natural history.¹⁸¹ Briefly, a model of colorectal cancer natural history was programmed and prior distributions were specified for all model parameters. Markov Chain Monte Carlo (MCMC) methods were used to estimate model parameters using data from multiple sources. For parameters that were unidentifiable using available data, informative prior distributions were specified; these distributions appropriately accounted for parameter uncertainty (as opposed to fixing the parameters to arbitrary values).

^c For example, this is true when the objective function is differentiable and the gradient of the objective function with respect to the calibration parameters is not a function of the remaining (other) parameters in the mathematical model, and the gradient of the objective function with respect to the other parameters is not a function of the calibration parameters.

The finite sample size performance of the proposed methodology was assessed in a simulation study, which demonstrated that the proposed method was an unbiased estimator for parameters for which data were available.

Nonetheless, jointly performing calibration and estimation of model parameters does not eliminate problems of identifiability: model parameters for which there is only limited (e.g., indirect or partial) or no information are effectively unidentifiable.¹⁸² Their posterior distribution is determined by the prior distribution chosen for them. In addition, in complex models, identifiability is hard to assess by just examining the model equations or inspecting the posterior distributions it produces. Instead, quantitative assessment is necessary. In the above-mentioned colorectal cancer microsimulation study,¹⁸¹ informative prior distributions were specified for unidentifiable model parameters and the model diagnostics proposed by Garrett & Zeger (2000) were used to assess identifiability via overlap statistics.¹⁸³ The utility of this approach was also demonstrated in the simulation study.¹⁸¹

Examples of calibration as estimation

In addition to Rutter (2009),¹⁸¹ other examples of using Bayesian methods for model calibration, validation, and parameter estimation exist, both for healthcare and non-healthcare decision and simulation models. These studies vary in their complexity, the number of data sources and the amount of information available for model development and evaluation.^{184;185} Jackson (2013) and Whyte (2011) provide tutorials on using Bayesian evidence synthesis methods and provide code and data to reproduce the analyses.^{184;185}

Conclusions

This chapter provides an overview of the state-of-the science on model validation and calibration for healthcare models. It appears that in healthcare, methodological research on the calibration and validation of decision and simulation models has been limited to case-studies applying a small number of alternative approaches to a small number of models. Because such case studies produce results that are applicable to these particular models, and address only a small part of the complex and multifaceted methodological decisions that modelers make, we believe that there is need for further research on validation and verification methods.

Based on our review of the literature and discussions with the stakeholders (described in *Chapter 1*), we have identified the following candidate areas for future research, with a focus on areas that may be of interest to the Effective Health Care Program:

- Consideration should be given to the *development of reference models* to facilitate the use of validated decision and simulation models as adjuncts to systematic reviews.^{186;187} Because model validation and calibration are time consuming activities and because systematic reviews need to be prepared in a timely fashion, the use of modeling in systematic reviews could be facilitated by developing and validating reference models for high-impact conditions (e.g., as has been done in CISNET).¹⁸⁸ Such conditions could be selected among AHRQ's priority areas, by taking into consideration the potential value of using models to supplement reviews of published evidence in each area.
- Further research is needed for the *development, validation, and calibration of complex models that incorporate evidence from multiple sources*. Systematic reviews (e.g., comparative effectiveness reviews prepared by Evidence-based Practice Centers) often retrieve evidence that is *flawed* (as indicated by risk of bias assessments), *indirect* (e.g.,

addressing laboratory surrogates instead of clinical outcomes), *incomplete* (e.g., with missing data), and *conflicting* (clinically and methodologically heterogeneous). Under these conditions “global subjective assessments” of the evidence are prone to error (and bias).¹⁷⁴ Modeling can address these problems by synthesizing evidence in a statistically valid way and allowing a formal assessment of consistency, while making all assumptions explicit.

- Research is needed to determine “best practices” for validating and calibrating models that are intended for use across different settings and patient populations.^{66;189} Such methods would rely on developing criteria for *formalizing judgments on the adequacy of the validation process* (especially external, prospective, and predictive validation).
- Given the importance of cross-model validation (especially in the absence of relevant empirical data) and the increasing availability of models addressing the similar research questions further research is needed to *explore how discrepancies among models relate to the models’ potential for being prospectively and externally validated (against data)*.
- Methodological work is also needed to *identify optimal methods for communicating (e.g., visualizing) the validation and calibration methods used in complex models*. Such research is necessary for presenting complex models to applied modelers and – more importantly – lay “consumers” of decision and simulation model results.

In summary, model validation and calibration are fundamental processes for establishing the credibility of decision and simulation models. “Confronting models with data” is an important component of establishing their validity and correct parameterization.¹⁰⁹ Ongoing progress in statistical, operational, and computational methods can provide modelers with an expanding toolkit for validating and calibrating models. However, current empirical research is limited to methodological appraisals or case-studies of alternative methods. Future research should advance our understanding of the theoretical basis of model evaluation and use comprehensive simulation methods to compare alternative approaches.

References

- (1) Mandelblatt JS, Cronin KA, Bailey S et al. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. *2009*;151:738-747.
- (2) Kuntz K, Sainfort F, Butler M, Taylor B, Kulasingam S, Gregory S et al. Decision and Simulation Modeling in Systematic Reviews. Methods Research Report. (Prepared by the University of Minnesota Evidence-based Practice Center under Contract No. 290-2007-10064-I.) AHRQ Publication No. 11(13)-EHC037-EF. 2013. Rockville, MD, Agency for Healthcare Research and Quality. <http://www.effectivehealthcare.ahrq.gov/tools-and-resources/researcher-resources/>
- (3) Ades AE, Claxton K, Sculpher M. Evidence synthesis, parameter correlation and probabilistic sensitivity analysis. *2006*;15:373-381.
- (4) Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *1999*;18:341-364.
- (5) Claxton K, Ginnelly L, Sculpher M, Philips Z, Palmer S. A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme. *2004*;8:1-103.
- (6) Philips Z, Ginnelly L, Sculpher M et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess* *2004*;8:iii-xi, 1.
- (7) Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. *2010*;11:55.
- (8) Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC. The role of cost-effectiveness analysis in health and medicine. Panel on Cost-Effectiveness in Health and Medicine. *JAMA* *1996*;276:1172-1177.
- (9) Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models. A suggested framework and example of application. *2000*;17:461-477.
- (10) Siegel JE, Weinstein MC, Russell LB, Gold MR. Recommendations for reporting cost-effectiveness analyses. Panel on Cost-Effectiveness in Health and Medicine. *JAMA* *1996*;276:1339-1341.
- (11) Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. *JAMA* *1996*;276:1253-1258.

- (12) Concannon TW, Meissner P, Grunbaum JA et al. A new taxonomy for stakeholder engagement in patient-centered outcomes research. *J Gen Intern Med* 2012;27:985-991.
- (13) Economic analysis of health care technology. A report on principles. Task Force on Principles for Economic Analysis of Health Care Technology. *Ann Intern Med* 1995;123:61-70.
- (14) Decision analytic modelling in the economic evaluation of health technologies. A consensus statement. Consensus Conference on Guidelines on Economic Modelling in Health Technology Assessment. *Pharmacoeconomics* 2000;17:443-444.
- (15) Andronis L, Barton P, Bryan S. Sensitivity analysis in economic evaluation: an audit of NICE current practice and a review of its use and value in decision-making. *Health Technol Assess* 2009;13:iii, ix-61.
- (16) Bae EY, Lee EK. Pharmacoeconomic guidelines and their implementation in the positive list system in South Korea. *Value Health* 2009;12 Suppl 3:S36-S41.
- (17) Boulenger S, Nixon J, Drummond M, Ulmann P, Rice S, de PG. Can economic evaluations be made more transferable? *Eur J Health Econ* 2005;6:334-346.
- (18) Brennan A, Chick SE, Davies R. A taxonomy of model structures for economic evaluation of health technologies. *Health Econ* 2006;15:1295-1310.
- (19) Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--1. *Value Health* 2012;15:796-803.
- (20) Chilcott J, Tappenden P, Rawdin A et al. Avoiding and identifying errors in health technology assessment models: qualitative study and methodological review. *Health Technol Assess* 2010;14:iii-xii, 1.
- (21) Chiou CF, Hay JW, Wallace JF et al. Development and validation of a grading system for the quality of cost-effectiveness studies. *Med Care* 2003;41:32-44.
- (22) Cleemput I, van WP, Huybrechts M, Vrijens F. Belgian methodological guidelines for pharmacoeconomic evaluations: toward standardization of drug reimbursement requests. *Value Health* 2009;12:441-449.
- (23) Clemens K, Townsend R, Luscombe F, Mauskopf J, Osterhaus J, Bobula J. Methodological and conduct principles for pharmacoeconomic research. Pharmaceutical Research and Manufacturers of America. *Pharmacoeconomics* 1995;8:169-174.
- (24) Colmenero F, Sullivan SD, Palmer JA et al. Quality of clinical and economic evidence in dossier formulary submissions. *Am J Manag Care* 2007;13:401-407.

- (25) Davalos ME, French MT, Burdick AE, Simmons SC. Economic evaluation of telemedicine: review of the literature and research guidelines for benefit-cost analysis. *Telemed J E Health* 2009;15:933-948.
- (26) Detsky AS. Guidelines for economic analysis of pharmaceutical products: a draft document for Ontario and Canada. *Pharmacoeconomics* 1993;3:354-361.
- (27) Drummond M, Brandt A, Luce B, Rovira J. Standardizing methodologies for economic evaluation in health care. Practice, problems, and potential. *Int J Technol Assess Health Care* 1993;9:26-36.
- (28) Drummond M, Sculpher M. Common methodological flaws in economic evaluations. *Med Care* 2005;43:5-14.
- (29) Drummond M, Manca A, Sculpher M. Increasing the generalizability of economic evaluations: recommendations for the design, analysis, and reporting of studies. *Int J Technol Assess Health Care* 2005;21:165-171.
- (30) Drummond M, Barbieri M, Cook J et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. *Value Health* 2009;12:409-418.
- (31) Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. *BMJ* 1996;313:275-283.
- (32) Evers S, Goossens M, de VH, van TM, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. *Int J Technol Assess Health Care* 2005;21:240-245.
- (33) Fry RN, Avey SG, Sullivan SD. The Academy of Managed Care Pharmacy Format for Formulary Submissions: an evolving standard--a Foundation for Managed Care Pharmacy Task Force report. *Value Health* 2003;6:505-521.
- (34) Garattini L, Grilli R, Scopelliti D, Mantovani L. A proposal for Italian guidelines in pharmacoeconomics The Mario Negri Institute Centre for Health Economics. *Pharmacoeconomics* 1995;7:1-6.
- (35) Gartlehner G, West SL, Mansfield AJ et al. Clinical heterogeneity in systematic reviews and health technology assessments: synthesis of guidance documents and the literature. *Int J Technol Assess Health Care* 2012;28:36-43.
- (36) Glennie JL, Torrance GW, Baladi JF et al. The revised Canadian Guidelines for the Economic Evaluation of Pharmaceuticals. *Pharmacoeconomics* 1999;15:459-468.
- (37) Goldhaber-Fiebert JD, Stout NK, Goldie SJ. Empirically evaluating decision-analytic models. *Value Health* 2010;13:667-674.

- (38) Graf von der Schulenburg JM, Greiner W, Jost F et al. German recommendations on health economic evaluation: third and updated version of the Hanover Consensus. *Value Health* 2008;11:539-544.
- (39) Grutters JP, Seferina SC, Tjan-Heijnen VC, van Kampen RJ, Goettsch WG, Joore MA. Bridging trial and decision: a checklist to frame health technology assessments for resource allocation decisions. *Value Health* 2011;14:777-784.
- (40) Hay J, Jackson J. Panel 2: methodological issues in conducting pharmacoeconomic evaluations--modeling studies. *Value Health* 1999;2:78-81.
- (41) Hoomans T, Severens JL, van der RN, Delwel GO. Methodological quality of economic evaluations of new pharmaceuticals in The Netherlands. *Pharmacoeconomics* 2012;30:219-227.
- (42) Karnon J, Brennan A, Akehurst R. A critique and impact analysis of decision modeling assumptions. *Med Decis Making* 2007;27:491-499.
- (43) Karnon J, Goyder E, Tappenden P et al. A review and critique of modelling in prioritising and designing screening programmes. *Health Technol Assess* 2007;11:iii-xi, 1.
- (44) Kolasa K, Dziomdziora M, Fajutrao L. What aspects of the health technology assessment process recommended by international health technology assessment agencies received the most attention in Poland in 2008? *Int J Technol Assess Health Care* 2011;27:84-94.
- (45) Liberati A, Sheldon TA, Banta HD. EUR-ASSESS Project Subgroup report on Methodology. Methodological guidance for the conduct of health technology assessment. *Int J Technol Assess Health Care* 1997;13:186-219.
- (46) Lopez-Bastida J, Oliva J, Antonanzas F et al. Spanish recommendations on economic evaluation of health technologies. *Eur J Health Econ* 2010;11:513-520.
- (47) Lovatt B. The United Kingdom guidelines for the economic evaluation of medicines. *Med Care* 1996;34:DS179-DS181.
- (48) Luce BR, Simpson K. Methods of cost-effectiveness analysis: areas of consensus and debate. *Clin Ther* 1995;17:109-125.
- (49) Mason J. The generalisability of pharmacoeconomic studies. *Pharmacoeconomics* 1997;11:503-514.
- (50) McCabe C, Dixon S. Testing the validity of cost-effectiveness models. *Pharmacoeconomics* 2000;17:501-513.
- (51) McGhan WF, Al M, Doshi JA, Kamae I, Marx SE, Rindress D. The ISPOR Good Practices for Quality Improvement of Cost-Effectiveness Research Task Force Report. *Value Health* 2009;12:1086-1099.

- (52) Menon D, Schubert F, Torrance GW. Canada's new guidelines for the economic evaluation of pharmaceuticals. *Med Care* 1996;34:DS77-DS86.
- (53) Mullahy J. What you don't know can't hurt you? Statistical issues and standards for medical technology evaluation. *Med Care* 1996;34:DS124-DS135.
- (54) Mullins CD, Ogilvie S. Emerging standardization in pharmacoeconomics. *Clin Ther* 1998;20:1194-1202.
- (55) Mullins CD, Wang J. Pharmacy benefit management: enhancing the applicability of pharmacoeconomics for optimal decision making. *Pharmacoeconomics* 2002;20:9-21.
- (56) Murray CJ, Evans DB, Acharya A, Baltussen RM. Development of WHO guidelines on generalized cost-effectiveness analysis. *Health Econ* 2000;9:235-251.
- (57) Blackmore CC, Magid DJ. Methodologic evaluation of the radiology cost-effectiveness literature. *Radiology* 1997;203:87-91.
- (58) Canadian Coordinating Office for Health Technology Assessment. Guidelines for economic evaluation of pharmaceuticals (Brief record). 2013.
- (59) Neumann PJ, Stone PW, Chapman RH, Sandberg EA, Bell CM. The quality of reporting in published cost-utility analyses, 1976-1997. *Ann Intern Med* 2000;132:964-972.
- (60) Nuijten MJ, Pronk MH, Brorens MJ et al. Reporting format for economic evaluation. Part II: Focus on modelling studies. *Pharmacoeconomics* 1998;14:259-268.
- (61) Olson BM, Armstrong EP, Grizzle AJ, Nichter MA. Industry's perception of presenting pharmacoeconomic models to managed care organizations. *J Manag Care Pharm* 2003;9:159-167.
- (62) Paisley S. Classification of evidence in decision-analytic models of cost-effectiveness: a content analysis of published reports. *Int J Technol Assess Health Care* 2010;26:458-462.
- (63) Ramsey S, Willke R, Briggs A et al. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report. *Value Health* 2005;8:521-533.
- (64) Roberts M, Russell LB, Paltiel AD, Chambers M, McEwan P, Krahn M. Conceptualizing a model: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--2. *Value Health* 2012;15:804-811.
- (65) Sassi F, McKee M, Roberts JA. Economic evaluation of diagnostic technology. Methodological challenges and viable solutions. *Int J Technol Assess Health Care* 1997;13:613-630.
- (66) Sculpher MJ, Pang FS, Manca A et al. Generalisability in economic evaluation studies in healthcare: a review and case studies. *Health Technol Assess* 2004;8:iii-192.

- (67) Severens JL, van der Wilt GJ. Economic evaluation of diagnostic tests. A review of published studies. *Int J Technol Assess Health Care* 1999;15:480-496.
- (68) Siegel JE, Torrance GW, Russell LB, Luce BR, Weinstein MC, Gold MR. Guidelines for pharmacoeconomic studies. Recommendations from the panel on cost effectiveness in health and medicine. Panel on cost Effectiveness in Health and Medicine. *Pharmacoeconomics* 1997;11:159-168.
- (69) Sonnenberg FA, Roberts MS, Tsevat J, Wong JB, Barry M, Kent DL. Toward a peer review process for medical decision analysis models. *Med Care* 1994;32:JS52-JS64.
- (70) Soto J. Health economic evaluations using decision analytic modeling. Principles and practices--utilization of a checklist to their development and appraisal. *Int J Technol Assess Health Care* 2002;18:94-111.
- (71) Taylor RS, Elston J. The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK Health Technology Assessment reports. *Health Technol Assess* 2009;13:iii, ix-50.
- (72) 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:E22-E29.
- (73) Udvarhelyi IS, Colditz GA, Rai A, Epstein AM. Cost-effectiveness and cost-benefit analyses in the medical literature. Are the methods being used correctly? *Ann Intern Med* 1992;116:238-244.
- (74) Ungar WJ, Santos MT. The Pediatric Quality Appraisal Questionnaire: an instrument for evaluation of the pediatric health economics literature. *Value Health* 2003;6:584-594.
- (75) Vegter S, Boersma C, Rozenbaum M, Wilffert B, Navis G, Postma MJ. Pharmacoeconomic evaluations of pharmacogenetic and genomic screening programmes: a systematic review on content and adherence to guidelines. *Pharmacoeconomics* 2008;26:569-587.
- (76) von der SJ, Vauth C, Mittendorf T, Greiner W. Methods for determining cost-benefit ratios for pharmaceuticals in Germany. *Eur J Health Econ* 2007;8 Suppl 1:S5-31.
- (77) 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.
- (78) Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. *Med Decis Making* 2012;32:733-743.
- (79) Husereau D, Drummond M, Petrou S et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *BMJ* 2013;346:f1049.

- (80) Gold M. Panel on cost-effectiveness in health and medicine. 1996;34:DS197-DS199.
- (81) International Network of Agencies for Health Technology Assessment. Health Technology Assessment Resources. 2014.
- (82) Luce BR, Drummond M, Jonsson B et al. EBM, HTA, and CER: clearing the confusion. *Milbank Q* 2010;88:256-276.
- (83) Ades AE, Cliffe S. Markov chain Monte Carlo estimation of a multiparameter decision model: consistency of evidence and the accurate assessment of uncertainty. 2002;22:359-371.
- (84) Mathes T, Jacobs E, Morfeld JC, Pieper D. Methods of international health technology assessment agencies for economic evaluations--a comparative analysis. *BMC Health Serv Res* 2013;13:371.
- (85) Agency for Health Technology Assessment in Poland (AHTApol/Poland). <http://www.aotm.gov.pl/index.php?id=397> [serial online] 2014.
- (86) Canadian Agency for Drugs and Technologies in Health (CADTH/Canada). <http://www.cadth.ca/en> [serial online] 2014.
- (87) Danish Centre for Health Technology Assessment (DACEHTA/Denmark). <http://www.sst.dk/English/DACEHTA.aspx> [serial online] 2014.
- (88) Health Information and Quality Authority (HIQA/Ireland). <http://www.hiqa.ie/> [serial online] 2014.
- (89) National Authority of Medicines and Health Products (INFARMED/Portugal). <http://www.infarmed.pt/portal/page/portal/INFARMED/ENGLISH> [serial online] 2014.
- (90) Institute for Quality and Efficiency in Health Care (IQWiG/Germany). <http://www.sst.dk/English/DACEHTA.aspx> [serial online] 2014.
- (91) Belgian Federal Health Care Knowledge Centre (KCE/Belgium). <https://kce.fgov.be/> [serial online] 2014.
- (92) MAS (Medical Advisory Secretariat, within the Ontario Ministry of Health and Long-Term Care Health Strategies Division). http://www.health.gov.on.ca/english/providers/program/mas/tech/tech_mn.html [serial online] 2014.
- (93) Medical Services Advisory Committee (MASC/Australia). <http://www.sst.dk/English/DACEHTA.aspx> [serial online] 2014.
- (94) National Institute for Clinical Excellence (NICE/UK). <http://www.nice.org.uk/> [serial online] 2014.

- (95) Pharmaceutical Benefits Advisor Committee (PBAC, Australia). <http://www.health.gov.au/internet/main/publishing.nsf/Content/pbac-outcomes-info> [serial online] 2014.
- (96) Pharmaceutical Management Agency of New Zealand (PHARMAC/New Zealand). <http://www.pharmac.govt.nz/> [serial online] 2014.
- (97) AAZ (Agency for Quality and Accreditation in Health Care, Croatia). <http://www.aaz.hr/> [serial online] 2014.
- (98) National Institute for Clinical Excellence (NICE/UK). <http://www.hitap.net/en/splash> [serial online] 2014.
- (99) ICER (Institute for Clinical and Economic Review). <http://www.icer-review.org/> [serial online] 2014.
- (100) LBI (Ludwig Boltzmann Institute for Health Technology Assessment). <http://hta.lbg.ac.at/page/homepage> [serial online] 2014.
- (101) MHRA (Medicines and Healthcare Products Regulatory Agency). <http://www.mhra.gov.uk/index.htm/> [serial online] 2014.
- (102) NLM (National Library of Medicine). <http://www.nlm.nih.gov/> [serial online] 2014.
- (103) AHRQ (US Agency for Healthcare Research and Quality). <http://www.effectivehealthcare.ahrq.gov/tools-and-resources/researcher-resources/> [serial online] 2014.
- (104) CAST (Centre for Applied Health Services Research and Technology Assessment, University of Southern Denmark). http://www.sdu.dk/Om_SDU/Institutter_centre/CAST?sc_lang=en [serial online] 2014.
- (105) CDE (Center for Drug Evaluation). <http://www.cde.org.tw/English/Pages/e-default.aspx> [serial online] 2014.
- (107) Kuntz KM, Lansdorp-Vogelaar I, Rutter CM et al. A systematic comparison of microsimulation models of colorectal cancer: the role of assumptions about adenoma progression. *Med Decis Making* 2011;31:530-539.
- (108) Agency for Healthcare Research and Quality. The Effective Health Care Program Stakeholder Guide, Appendix C. 2011. Rockville, MD, Agency for Healthcare Research and Quality.
- (109) Cooper BS. Confronting models with data. *J Hosp Infect* 2007;65 Suppl 2:88-92.
- (110) Groves P, Kayyali B, Knott D, Van Kuiken S. The 'big data' revolution in healthcare. Center for US Health System Reform. Business Technology Office. 2013. McKinsey & Company.

- (111) Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System--a national resource for evidence development. *N Engl J Med* 2011;364:498-499.
- (112) Zechmeister-Koss I, Schnell-Inderst P, Zauner G. Appropriate Evidence Sources for Populating Decision Analytic Models within Health Technology Assessment (HTA): A Systematic Review of HTA Manuals and Health Economic Guidelines. *Med Decis Making* 2013.
- (113) Ades AE, Welton NJ, Caldwell D, Price M, Goubar A, Lu G. Multiparameter evidence synthesis in epidemiology and medical decision-making. *J Health Serv Res Policy* 2008;13 Suppl 3:12-22.
- (114) Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. Evidence synthesis for decision making 7: a reviewer's checklist. *Med Decis Making* 2013;33:679-691.
- (115) Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 1: introduction. *Med Decis Making* 2013;33:597-606.
- (116) Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 6: embedding evidence synthesis in probabilistic cost-effectiveness analysis. *Med Decis Making* 2013;33:671-678.
- (117) Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 5: the baseline natural history model. *Med Decis Making* 2013;33:657-670.
- (118) Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. *Med Decis Making* 2013;33:641-656.
- (119) Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: heterogeneity--subgroups, meta-regression, bias, and bias-adjustment. *Med Decis Making* 2013;33:618-640.
- (120) Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making* 2013;33:607-617.
- (121) ISPOR. Value in Health Guide for Authors. 2012;2012.
- (122) Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. 2007;146:450-453.
- (123) Peng RD. Reproducible research in computational science. 2011;334:1226-1227.
- (124) Weinstein MC. Recent developments in decision-analytic modelling for economic evaluation. 2006;24:1043-1053.

- (125) Morin A, Urban J, Adams PD et al. Research priorities. Shining light into black boxes. 2012;336:159-160.
- (126) Weinstein MC, Toy EL, Sandberg EA et al. Modeling for health care and other policy decisions: uses, roles, and validity. 2001;4:348-361.
- (127) Committee on Mathematical and Statistical Foundations of Verification VaUQ. Assessing the Realiability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Qualification. 2012. Washington, D.C., The National Academies Press.
- (128) Kim LG, Thompson SG. Uncertainty and validation of health economic decision models. 2010;19:43-55.
- (129) Berry DA, Cronin KA, Plevritis SK et al. Effect of screening and adjuvant therapy on mortality from breast cancer. 2005;353:1784-1792.
- (130) Moriasi D, Wilson B, Douglas-Mankin K, Arnold J, Gowda P. Hydrologic and water quality models: use, calibration, and validation. *Transactions of the ASABE* 2012;55:1241-1247.
- (131) Legates DR, McCabe GJ. Evaluating the use of "goodness of fit measures in hydrologic and hydroclimatic model validation.". *Water resources research* 1999;35:233-241.
- (132) Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans ASABE* 2007;50:885-900.
- (133) Rao C, Wu Y. On model selection. Institute of Mathematical Statistics Lecture Notes-Monograph Series 38, 1-57. 2001.
- (134) Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 6, 733-807. 1996.
- (135) Gelman A. A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-fit Testing*. *International Statistical Review* 2003;71:369-382.
- (136) Gardner A, Deardon R, Darlington G. Goodness-of-fit measures for individual-level models of infectious disease in a Bayesian framework. 2011;2:273-281.
- (137) Deuchert E, Brody S. Plausible and implausible parameters for mathematical modeling of nominal heterosexual HIV transmission. 2007;17:237-244.
- (138) Hur C, Hayeck TJ, Yeh JM et al. Development, calibration, and validation of a U.S. white male population-based simulation model of esophageal adenocarcinoma. 2010;5:e9483.

- (139) Ishida H, Wong JB, Hino K et al. Validating a Markov model of treatment for hepatitis C virus-related hepatocellular carcinoma. 2008;47:529-540.
- (140) Nijhuis RL, Stijnen T, Peeters A, Witteman JC, Hofman A, Hunink MG. Apparent and internal validity of a Monte Carlo-Markov model for cardiovascular disease in a cohort follow-up study. 2006;26:134-144.
- (141) Perreault S, Levinton C, Laurier C, Moride Y, Ste-Marie LG, Crott R. Validation of a decision model for preventive pharmacological strategies in postmenopausal women. 2005;20:89-101.
- (142) Sendi PP, Craig BA, Pfluger D, Gafni A, Bucher HC. Systematic validation of disease models for pharmacoeconomic evaluations. Swiss HIV Cohort Study. *Journal of Evaluation in Clinical Practice* 1999;5:283-295.
- (143) Willis M, Asseburg C, He J. Validation of economic and health outcomes simulation model of type 2 diabetes mellitus (ECHO-T2DM). 2013;16:1007-1021.
- (144) Vanni T, Karnon J, Madan J et al. Calibrating models in economic evaluation: a seven-step approach. 2011;29:35-49.
- (145) Gourieroux C, Monfort A, Trognon A. Pseudo maximum likelihood methods: Theory. *Econometrica: Journal of the Econometric Society* 1984;681-700.
- (146) Gong G, Samaniego FJ. Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics* 1981;861-869.
- (147) Boyd SP, Vandenberghe L. *Convex optimization*. Cambridge university press, 2004.
- (148) Bertsimas D, Farias VF, Trichakis N. On the efficiency-fairness trade-off. *Management Science* 2012;58:2234-2250.
- (149) Nonlinear Programming. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, California: University of California Press, 2008.
- (150) Vapnik V. Statistical learning theory. 1998. Wiley, New York.
- (151) Stout NK, Knudsen AB, Kong CY, McMahon PM, Gazelle GS. Calibration methods used in cancer simulation models and suggested reporting guidelines. *Pharmacoeconomics* 2009;27:533-545.
- (152) Earnshaw SR, Wilson M, Mauskopf J, Joshi AV. Model-based cost-effectiveness analyses for the treatment of acute stroke events: a review and summary of challenges. 2009;12:507-520.
- (153) Rochau U, Schwarzer R, Jahn B et al. Systematic assessment of decision-analytic models for chronic myeloid leukemia. 2014;12:103-115.

- (154) Abuelezam NN, Rough K, Seage GR, III. Individual-based simulation models of HIV transmission: reporting quality and recommendations. 2013;8:e75624.
- (155) Bolin K. Economic evaluation of smoking-cessation therapies: a critical and systematic review of simulation models. 2012;30:551-564.
- (156) Punyacharoensin N, Edmunds WJ, De AD, White RG. Mathematical models for the study of HIV spread and control amongst men who have sex with men. 2011;26:695-709.
- (157) Goehler A, Geisler BP, Manne JM et al. Decision-analytic models to simulate health outcomes and costs in heart failure: a systematic review. 2011;29:753-769.
- (158) Haji Ali AH, Gray J, Karnon J. Model performance evaluation (validation and calibration) in model-based studies of therapeutic interventions for cardiovascular diseases : a review and suggested reporting framework. 2013;11:85-93.
- (159) Dams J, Bornschein B, Reese JP et al. Modelling the cost effectiveness of treatments for Parkinson's disease: a methodological review. 2011;29:1025-1049.
- (160) Ferdinands JM, Mannino DM. Obstructive lung disease models: what is valid? 2008;5:382-393.
- (161) Unal B, Capewell S, Critchley JA. Coronary heart disease policy models: a systematic review. 2006;6:213.
- (162) Siebert U, Bornschein B, Walbert T, Dodel RC. Systematic assessment of decision models in Parkinson's disease. 2004;7:610-626.
- (163) McMahon PM, Kong CY, Johnson BE et al. Chapter 9: The MGH-HMS lung cancer policy model: tobacco control versus screening. 2012;32 Suppl 1:S117-S124.
- (164) Kong CY, McMahon PM, Gazelle GS. Calibration of disease simulation model using an engineering approach. 2009;12:521-529.
- (165) Kohli M, Ferko N, Martin A et al. Estimating the long-term impact of a prophylactic human papillomavirus 16/18 vaccine on the burden of cervical cancer in the UK. 2007;96:143-150.
- (166) Gulati P, Tripathi RP, Jena A. Magnetic resonance imaging in brain lesions. 1990;27:1327-1332.
- (167) Karnon J, Vanni T. Calibrating models in economic evaluation: a comparison of alternative measures of goodness of fit, parameter search strategies and convergence criteria. 2011;29:51-62.
- (168) Karnon J, Delea T, Barghout V. Cost utility analysis of early adjuvant letrozole or anastrozole versus tamoxifen in postmenopausal women with early invasive breast cancer: the UK perspective. 2008;9:171-183.

- (169) Taylor DC, Pawar V, Kruzikas DT, Gilmore KE, Sanon M, Weinstein MC. Incorporating calibrated model parameters into sensitivity analyses: deterministic and probabilistic approaches. 2012;30:119-126.
- (170) Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine* 2003;22:3687-3709.
- (171) Eddy DM. The confidence profile method: a Bayesian method for assessing health technologies. *Operations Research* 1989;37:210-228.
- (172) Eddy DM, Hasselblad V, Shachter R. A Bayesian method for synthesizing evidence. The Confidence Profile Method. *International Journal of Technology Assessment in Health Care* 1990;6:31-55.
- (173) Eddy DM, Hasselblad V, Shachter R. An introduction to a Bayesian method for meta-analysis: The confidence profile method. *Medical Decision Making* 1990;10:15-23.
- (174) Eddy D, Shachter R. *Meta-Analysis by the Confidence Profile Method*. Academic Press, 1992.
- (175) Droitcour J, Silberman G, Chelimsky E. A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *International Journal of Technology Assessment in Health Care* 1993;9:440-449.
- (176) Kaizar EE. Estimating treatment effect via simple cross design synthesis. *Statistics in Medicine* 2011;30:2986-3009.
- (177) Nixon RM, Duffy SW. Cross-issue synthesis: potential application to breast cancer, tamoxifen and genetic susceptibility. *Journal of Cancer Epidemiology & Prevention* 2002;7:205-212.
- (178) Wald NJ, Morris JK. Teleoanalysis: combining data from different types of study. *BMJ* 2003;327:616-618.
- (179) Welton NJ, Ades AE. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. 2005;25:633-645.
- (180) Marshall EC, Spiegelhalter DJ. Approximate cross-validators predictive checks in disease mapping models. *Statistics in Medicine* 2003;22:1649-1660.
- (181) Rutter CM, Miglioretti DL, Savarino JE. Bayesian Calibration of Microsimulation Models. 2009;104:1338-1350.
- (182) Basu A. Identifiability. In: Kotz S, Johnson N, eds. *Encyclopedia of Statistical Sciences*. Wiley Interscience: 2006;2-6.

- (183) Garrett ES, Zeger SL. Latent class model diagnosis. *Biometrics* 2000;56:1055-1067.
- (184) Jackson CH, Jit M, Sharples LD, De AD. Calibration of Complex Models through Bayesian Evidence Synthesis: A Demonstration and Tutorial. 2013.
- (185) Whyte S, Walsh C, Chilcott J. Bayesian calibration of a natural history model with application to a population model for colorectal cancer. *Medical Decision Making* 2011;31:625-641.
- (186) Afzali HH, Karnon J, Merlin T. Improving the accuracy and comparability of model-based economic evaluations of health technologies for reimbursement decisions: a methodological framework for the development of reference models. *Medical Decision Making* 2013;33:325-332.
- (187) Haji Ali AH, Karnon J. Addressing the challenge for well informed and consistent reimbursement decisions: the case for reference models. *PharmacoEconomics* 2011;29:823-825.
- (188) Cancer Intervention and Surveillance Modeling Network. <http://cisnet.cancer.gov/> [serial online] 2014.
- (189) Mason JM, Mason AR. The generalisability of pharmacoeconomic studies: issues and challenges ahead. *PharmacoEconomics* 2006;24:937-945.