Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force–7

David M. Eddy, PhD, MD, William Hollingworth, PhD, J. Jaime Caro, MDCM, FRCPC, FACP, Joel Tsevat, MD, MPH, Kathryn M. McDonald, MM, John B. Wong, MD, On Behalf of the ISPOR-SMDM Modeling Good Research Practices Task Force

Trust and confidence are critical to the success of health care models. There are two main methods for achieving this: transparency (people can see how the model is built) and validation (how well it reproduces reality). This report describes recommendations for achieving transparency and validation, developed by a task force appointed by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the Society for Medical Decision Making (SMDM). Recommendations were developed iteratively by the authors. A nontechnical description should be made available to anyone—including model type and intended applications; funding sources; structure; inputs, outputs, other components that determine function, and their relationships; data sources; validation methods and results; and limitations. Technical documentation, written in sufficient detail to enable a reader with necessary expertise to evaluate the model and potentially reproduce it, should be made available openly or under agreements that protect intellectual property, at the discretion of the modelers. Validation involves face validity (wherein experts evaluate model structure, data sources, assumptions, and results), verification or internal validity (check accuracy of coding), cross validity (comparison of results with other models analyzing same problem), external validity (comparing model results to real-world results), and predictive validity (comparing model results with prospectively observed events). The last two are the strongest form of validation. Each section of this paper contains a number of recommendations that were iterated among the authors, as well as the wider modeling task force jointly set up by the International Society for Pharmacoeconomics and Outcomes Research and the Society for Medical Decision Making. ***Key words:*** *modeling; transparency; validation; good practices; simulation; decision sciences.* ***(Med Decis Making 2012;32:733–743)***

A new Good Research Practices in Modeling Task Force was constituted by the ISPOR Board of Directors in 2010, and the Society for Medical Decision Making was invited to join the effort. This paper, along with six others,[1-6] is part of a series commissioned by the Task Force.

Received 2 March 2012 from Archimedes, Inc., San Francisco, CA (DE); School of Social and Community Medicine, University of Bristol, Bristol, UK (WH); United BioSource Corporation and McGill University, Montreal, Canada (JJC); University of Cincinnati, College of Medicine and Cincinnati Veterans Affairs Medical Center, Cincinnati, OH (JT); Center for Health Policy/Center for Primary Care and Outcomes Research, Stanford, CA (KM); Tufts University School of Medicine, Boston, MA (JW). Revision accepted for publication 20 June 2012.

Address correspondence to David Eddy, Archimedes, Inc., 201 Mission St. Suite 2900, San Francisco, CA 94105 USA; e-mail: david.eddy@archimedesmodel.com.

DOI: 10.1177/0272989X12454579

INTRODUCTION

The purpose of health care models is to provide decision makers with quantitative information about the consequences of the options being considered. For a model to be useful for this purpose, decision makers need confidence in the model's results.

Related Materials

For more information on the ISPOR-SMDM Task Force, visit the website at http://www.ohsu.edu/epc/mdm/modeling.cfm. See ''Modeling Good Research Practices—Overview, Issues, and Preferred Practices: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force,'' by J. Jaime Caro, Andrew H. Briggs, Uwe Siebert, and Karen M. Kuntz, published in this issue on pages 667-677, for an overview of the series. See ''Transparently, with Validation,'' by Nancy Neil, published in this issue on pages 660-662, and ''Transparency and Reproducible Research in Modeling: Why We Need It and How to Get There,'' by Crystal M. Smith-Spangler, published in this issue on pages 663-666, for additional information.

Specifically, they need to know how accurately the model predicts the outcomes of interest, and they need to account for that information when deciding how to use the model results.

Modelers can impart such confidence and enhance model credibility in two main ways:

transparency—clearly describing the model structure, equations, parameter values, and assumptions to enable interested parties to understand the model; and
validation—subjecting it to tests, such as comparing model results with events observed in reality.[7–14]

Some health care models are intended to be ''general'' or ''multiapplication'' in the sense that with appropriate modifications, they can address a range of problems.[15,16] For example, an ''HIV model'' could be used repeatedly to address different questions relating to that condition.[17–21] Other models are built for single, specific applications and are not intended to be reused.[22] For instance, a model may be built with the sole purpose of extrapolating the results of a trial of an implantable cardioverter defibrillator to determine if it becomes cost-effective over the lifetime of patients.[23] Some models that are initially built for a single application may later be expanded to address others. The methods described in this paper apply to both types of models. For a multiapplication model, transparency, validation, and reporting are ongoing processes. The multiapplication model is described (transparency)[24] and validated,[25] and the descriptions and validations are continuously updated as science and the model evolve.[26] In addition, each instantiation of the model is described, validated, and reported as each application is done.[27] For a single-application model, its description and validation and the reporting of its application are typically conducted at one time, although there may be additional validations after initial use, particularly if problems are found.

Our objective is to describe practices that we consider to be ''best'' in the sense of providing potential users of a model with the information necessary to determine their confidence in the results and, hence, their application of the model's results. Every model today should be able to achieve the best practices that we recommend for transparency. We recognize, however, that not all models will be able to achieve all the recommended best practices for validation. Rather than establish minimum quality standards, we have described optimal practices that all models should strive toward. For all models, their developers should describe their process for conducting validations and the level of validation their model achieved. These recommendations are particularly important in light of high-profile examples of scientific misconduct and fraudulent research published in leading scientific journals, leading to increasing emphasis on transparency and ''shining a light on black boxes.''[28–34]

TRANSPARENCY

Transparency refers to the extent to which interested parties can review a model's structure, equations, parameter values, and assumptions. It does not refer to the formulation, conduct, or results of a particular analysis. Transparency serves two purposes: 1) to provide a nonquantitative description of the model to readers who want to understand in a general way how a model works and 2) to provide technical information to readers[1] who want to evaluate a model at higher level of mathematical and programming detail and possibly replicate it. Taken together, the intention is to provide sufficient information to enable the full spectrum of readers to understand a model's accuracy, limitations, and potential applications at a level appropriate to their expertise and needs.

Nontechnical Documentation

Nontechnical documentation should be accessible to any interested reader.[35–37] It should include descriptions of

1. the model and its purpose;
2. the types of applications that it is designed to address (e.g., forecasting of short-term costs, cost-effectiveness analysis);
3. sources of funding and their role;
4. structure (e.g., graphical representation of the variables and their relationships);
5. components that define it and determine its performance;
6. inputs, outputs, other parameters;
7. equations and their sources;
8. how the data sources were identified and selected;
9. model validation and summary of results;
10. methods for customizing to specific applications and settings;
11. effects of uncertainty;
12. main limitations for its intended applications;
13. examples of actual equations (optional); and
14. reference to the model's technical documentation

The nontechnical documentation provides an overview of the model and what it does, but it may not contain sufficient information to enable readers to replicate it.

Technical Documentation

Full technical transparency is achieved by providing documents that detail the model, including its structure,

components, equations, and computer code. The documentation should be sufficiently detailed to enable those with the necessary expertise and resources to reproduce the model. Provision of technical documentation is subject to some conditions and limitations:

1. Access should be provided in a way that enables protection of intellectual property. Building a model can require a significant investment in time and money; if those who make such investments had to give their models away without restriction, the incentives and resources to build and maintain complex models could disappear.
2. While not mandatory, an increasing number of journals request that authors state whether full technical documentation is available to readers and, if so, under what terms.[28,38] Technical documents may be placed in appendices or made accessible by other means.[28,29,31,39] Provision of such documentation is not without concerns that the context of the original analysis may be missing.[40]
3. Because most multiapplication models change over time—expanded and updated to incorporate new information and advances in health care technologies—technical documents should be updated periodically.
4. Equations and detailed structure will mean little to readers without the necessary technical background. Even with such information, reviewing a model can take considerable time. Furthermore, it is very difficult to understand how accurate a model is by simply examining its equations. Even if the equations appear to be valid in a mathematical sense and the parameters appear to be estimated using appropriate sources and methods, it is virtually impossible for anyone to determine a model's accuracy by ''running'' it in their heads. Providing the code does not solve this problem unless the reader has the time and resources to actually implement it, which can be very difficult for large models or models that require advanced computing methods (e.g., distributed computing). Provision of code in this way would also threaten protection of intellectual property. Some of these limitations can be addressed by giving readers access to the model or to a version applicable to a particular analysis. Even enabling readers to specify inputs and receive outputs of a model without releasing a full copy of it can provide useful information about how the model functions. Thus, if feasible, modelers should give readers access to the model itself or parts applicable to a particular analysis. Having said this, it is important to note that providing such access can be very expensive, including the cost to build the copy and interfaces and the support to ensure that the model is used and interpreted accurately. Modelers who make working copies accessible to readers should be credited for providing an exceptionally high level of transparency, but failure to do this should not imply any failure on the part of the modeler and should not prejudice an evaluation of the model.

Public Versus Confidential Documentation

To address the conflicting needs of transparency versus feasibility and intellectual property, information put into the public domain without restriction (''public documentation'') should be distinguished from information made available under agreements that protect intellectual property (''confidential documentation''). For public documentation, the nontechnical description should be available to all who ask. In addition, at their discretion, modelers can choose to make technical documentation or a working copy of the model publically available. Regarding confidential documentation, modelers should provide full technical documentation (along with access to a working copy) to readers designated either by a journal reviewing a paper or by an organization to which the model is provided for decision making, under agreements protecting intellectual property.[2] Providing readers with technical information under conditions of confidentiality is consistent with published requirements for the review of models.[41] Given the documentation's size and technical nature the need to protect intellectual property and because journals can gain full access to all documentation during the review process, journals should not require that it be included in the published report of an analysis.

Best practices

VII-1 Every model should have nontechnical documentation that should be freely accessible to any interested reader. At a minimum, it should describe in nontechnical terms the type of model and intended applications; funding sources; model structure; inputs, outputs, other components that determine the model's function, and their relationships; data sources; validation methods and results; and limitations.

VII-2 Every model should have technical documentation, written in sufficient detail to enable a reader with the necessary expertise to evaluate the model and potentially reproduce it. The technical documentation should be made available openly or under agreements that protect intellectual property, at the discretion of the modelers.

Expectations

Even with these best practices, it will rarely be possible to make any model completely transparent to all

readers (e.g., given the need for advanced mathematical or computer science training). Lack of transparency to those who do not have the appropriate training or time to study a model does not imply that the model is necessarily flawed. It is also important to stress that transparency by itself does not imply a model is accurate. A model can be transparent but yield the wrong answer (e.g., the formula *distance* = *rate*/*time* is transparent but wrong). Conversely, a model can lack transparency for most readers but be correct. Ultimately, what matters is whether a model accurately predicts what occurs in reality. Thus, transparency and validation are inextricably linked, and both are required to help readers gain confidence in a model's results. Analogously, the equations used to convert CT scans into images are not transparent to most physicians, yet physicians use CT scans all the time. They are willing to do so because operations based on a CT scan showing a mass almost always find a mass. The key to developing confidence in a model is not just studying its structure, assumptions, equations, and code but assessing whether it accurately calculates the outcomes of interest. This is the role of validation.

VALIDATION

Importance of Validation

Validation is a set of methods for judging a model's accuracy in making relevant predictions. That information can be used by decision makers to determine the results' applicability to their decision. While transparency can help readers understand what a model does and how it does it, validation is the only way for readers to determine how well it does it.

Validation is vital for both multiple- and single-application models. For multiple-application models, a distinction should be made between validating in a general sense (e.g., a ''diabetes model'' or a ''heart disease model'') and validating for a specific application (e.g., the effect of a drug on glycemia in diabetes). For a single-application model, validation can be limited to that application.

It is not possible to specify criteria that a model must meet to be declared ''valid,'' as if validity were a property of the model that applies to all of its applications and uses for all time. Because a model can have different levels of validity for different applications, the concept of validation should apply to particular applications, not to the model itself. Second, the required degree of accuracy depends on the question. For example, much less accuracy is needed to inform ''Will this intervention increase or decrease costs?'' than to answer ''How much will this intervention cost?'' A third reason applies to multiapplication models—they can and should change over time to keep up with new science, technologies, and evidence. Thus, all validation types should be conducted in the context of specific applications, and their reports should include the intended applications.

No matter how many validations are done, there will inevitably be uncertainty about some aspects of a model. Sensitivity analysis can be used to explore how a model's results change upon variation in inputs,[6] but by itself, it does not evaluate how accurately a model simulates what occurs in reality. Sensitivity analysis is an important complement to validation but not a substitute for it.

Types of Validation

Five main types of validation are commonly described: face validity, verification (or internal validity), cross validity, external validity, and predictive validity. Face validity is the extent to which a model and its assumptions and applications correspond to current science and evidence, as judged by people who have expertise in the problem. Verification addresses whether the model's parts behave as intended and the model has been implemented correctly. Cross validation involves comparing a model with others and determining the extent to which they calculate similar results. In external validation, a model is used to simulate a real scenario, such as a clinical trial, and the predicted outcomes are compared with the real-world ones. Predictive validity involves using a model to forecast events and, after some time, comparing the forecasted outcomes to the actual ones. Each type of validation has methods, strengths, limitations, and best practices.

Face Validity

Four aspects are particularly important for face validity: model structure, data sources, problem formulation, and results. Face validity is subjective; people who have clinical expertise should evaluate how well each model component reflects their understanding of the pertinent medical science, available evidence, and the clinical or administrative question at issue.[42] Information about the model and supporting evidence is obtained from documentation provided by the modelers. Information about the problem formulation and results is obtained from the application's report.

Specific questions depend on the component being evaluated. For the structure, important questions are whether the model includes all aspects of reality considered important by experts and whether they are related in ways consistent with medical science; for evidence, whether the best available data sources were used; for problem formulation, whether the setting, population, interventions, outcomes, assumptions, and time horizons correspond to those of interest; for results, whether they match experts' expectations and, if not, whether the model can plausibly explain them. If perceived weaknesses exist in any of these aspects, the assessment should examine how well the authors have reported and explained the discrepancies and what potential effects they have on the results. For example, if a model omits an important risk factor, have the modelers described the direction and potential magnitude of any resulting bias?

Evaluation of face validity can occur in several ways. The group that developed the model can appeal to members of the modeling group, to people in the same organization who did not build the model, or to external consultants. Any readers can perform their own evaluation. Peer review typically includes an evaluation of face validity. Because face validity is subjective, the evaluators should have no stake in the problem at issue. Ideally, the structure, evidence, and problem formulation should be assessed without knowing the results. As presently practiced, peer review of a model for publication is insufficiently consistent to be relied on for determining face validity.

Strengths and limitations. Face validation helps to ensure that a model is constructed and used in accord with most current medical science and best available evidence. This process enhances credibility with experts and increases acceptance of results. Additionally, the evaluation can raise questions and force thinking that improves the model. If the results are counterintuitive but justified, exploring the causes can identify new hypotheses and stimulate additional data collection and research.

Face validation has several limitations, however. All models simplify reality, many to a very large extent. Thus, the structure may not be completely consistent with medical knowledge or beliefs and would not have face validity if strictly applied. For example, physicians know that representing a complex disease as a small number of discrete states is clinically unrealistic and that patients do not jump from one state to another at fixed time intervals as such occurs in state–transition models. Despite these simplifications, for properly selected problems, state–transition models can be sufficiently accurate to meet the needs.[3] It can be very difficult for readers to determine in their heads whether a model has been properly simplified, oversimplified, or undersimplified for a particular problem.

Another limitation is that current medical evidence is incomplete and that medical knowledge and beliefs can be wrong or can change. Insistence on expert agreement with all aspects of the model's structure at any particular time can build misconceptions into a model. For example, until recently, virtually all experts believed that raising HDL cholesterol would prevent cardiovascular events. Models excluding that presumed effect would have low face validity, yet a recent clinical trial contradicted the presumption.[43]

A third limitation is that there are no unambiguous criteria to apply to judgments about a model or its application. Lacking these, it is easy for anyone who has a stake to be swayed. Specifically, virtually all modelers would say that their model has face validity. Anyone with a stake in the results may have a bias toward accepting a model upon liking its results or rejecting it if not.

Best practices

VII-3 Validation should include an evaluation of face validity of a model's structure, evidence, problem formulation, and results. A description of the process used to evaluate face validity should be made available on request. To the greatest extent possible, evaluation of face validity should be made by people who have expertise in the problem area but are impartial and preferably blinded to the results of the analysis. If face validation raises questions, these issues should be discussed in the report.

Verification

Verification (also called internal validity, internal consistency, or technical validity[42,44–46]) is a type of validity that examines the extent to which the mathematical calculations are performed correctly and are consistent with the model's specifications. The methods depend on the model's complexity. There are two main steps: verifying the individual equations and verifying their accurate implementation in code. Equations and parameters should be validated against their sources. Coding accuracy should be checked using state-of-the-art quality assurance and control methods for software engineering.[45,46] Examples of techniques include the following: maintaining complete and up-to-date documentation of the code;

conducting structured ''walk-throughs'' in which the programmer explains the code to other people, who search for errors; verifying separate parts of a model one by one; double programming, in which sections of a model are programmed independently by two programmers; comparing a model's results with hand calculations; performing sensitivity analysis, extreme value analysis, and trace analysis (in which individual events and their timing are tracked); and identifying unnecessary detail that might increase the likelihood of errors. The choice of methods should be appropriate for a model's complexity.

Strengths and limitations. Verification helps to ensure that there are no unintentional computational errors, but it does not evaluate the accuracy of the model's structure or predictions. Parameters for the equations might be fitted using good data sources and technique, and the equations might be accurately coded, but the resulting model might still be inaccurate if the structure is poorly chosen. For example, if a question involving distance, rate, and time is set up as $D = \alpha + \beta 1R + \beta 2T$, instead of $D = R \times T$, the parameters α, $\beta 1$, and $\beta 2$ can be estimated properly and the equation can be coded correctly, but the results could be wrong, depending on the ranges of R and T. Verification will not identify such problems. Verification should also involve sensitivity analysis of all parameters, evaluating a broad range of input values to determine if the direction and magnitude of model outputs behave as expected.

Best practices

VII-4 Models should be subjected to rigorous verification. The methods should be described in the model's nontechnical documentation. Pertinent results of verification should be made available on request.

Cross Validation

Cross validation (also called external consistency, comparative modeling, external convergence testing, convergent validity, external consistency, model corroboration) is a method that involves examining different models that address the same problem and comparing their results.[9,47–50] The differences among the results and their causes are then examined.

Strengths and limitations. Confidence in a result is increased if similar results are calculated by models using different methods.[9,13] Comparisons across models can also be useful for methodological purposes. The meaningfulness of this type of validation depends on the degree to which the methods and data sources of the different models are independent. A high degree of dependency among models (e.g., using parameters from other models published earlier) reduces the value of cross validation. Alternative structures and assumptions, as with the seven independent breast cancer models of the Cancer Intervention and Surveillance Modeling Network, enhance the credibility of the cross validation.[48]

Best practices

VII-5 Modelers should search for modeling analyses of the same or similar problems and discuss insights gained from similarities and differences in results.

External Validation

External validation compares a model's results to actual event data. It involves simulating events that have occurred, such as those in a clinical trial, and examining how well the results correspond. For multiapplication models, external validations can be applied to the model in a general sense and to each application. They can also be applied to the model as a whole or to some components. It is important to perform multiple validations that crisscross the intended applications in the sense of involving a range of populations, interventions, outcomes, and time horizons. External validation can also be applied to model components, such as population creation, disease incidence (including effects of patient characteristics, risk factors, and behaviors), disease progression, care processes and behaviors, occurrence of clinical outcomes, and interventions and their effects (except for utilization and procedure outcomes, or financial costs—see limitations). Examples of component validation include using an epidemiological study to validate a model's incidence equations and using biomarker progression in a trial's control arm to validate physiologic equations. In contrast, simulation of an entire clinical trial tests several or all parts at once, and simulation of multiple trials tests the model's accuracy in calculating multiple outcomes in multiple populations treated with multiple interventions.

External validation and predictive validation are critical as they most closely correspond to the model's purpose—to help decision makers anticipate what will occur if they take certain actions. There are three main steps: identifying the data sources to simulate, conducting the simulation, and comparing results.

Identifying the data sources. Data sources must fulfill two requirements: 1) contain applicable data

and 2) be sufficiently described to enable replication of design (information about the setting, population, treatment protocols, follow-up protocols, and outcomes) and progression (any changes in the design or conduct of the study over the follow-up period). Examples of data sources include population statistics, epidemiologic studies, clinical trials, claims data, and electronic health records.

Sources can be either formal or informal. A formal data source is one intended for research purposes and includes explicit planning and description of study design, selection criteria, data gathering and recording methods, intervention protocols, follow-up protocols, outcome definitions, specified follow-up time, and methods used to aggregate results and report outcomes. An informal data source is intended primarily for other purposes (e.g., clinical records, claims data). The distinction is important because simulating a source is difficult without explicit planning and descriptions. Whether formal or informal, the quality of sources can vary widely.

A validation is dependent if the same source was used both to estimate model parameters and to validate the model. A validation is partially dependent if the source was used to build or calibrate part of a model, but that part by itself does not wholly determine the outcome to be validated. Thus, a source can be dependent for one part but ''partially dependent'' for another. A validation is independent if no information from the source was used to build the model. An independent validation is blinded if those performing the validation had no information about the outcomes in the source. For example, papers on the design and initial conditions of a trial can be used to try to predict its results. Even when a data source is unblinded, those conducting an independent validation should not allow information about the outcomes to influence the validation.

Calibration has been described as ensuring that the ''inputs and outputs are consistent with available data,''[13] which may be dependent, partially dependent, or independent. A common form of calibration involves constraining the possible values for unobservable model parameters by matching model output to external data, as in cancer simulation.[51] Calibration reporting should include the target, goodness-of-fit metric, search algorithm, acceptance criteria, and stopping rule.[51] Thus, dependent validations are closely related to calibration, where data are used to fit model parameters.

Ideally, the validation plan and sources should be chosen before the results are known. Sources should be identified via a formal search using established methods,[52–55] selecting those with settings, populations, interventions, and outcomes similar to those the model addresses and those with the best designs (e.g., large size, representative population, formal protocol, detailed reporting, and recency). The choices might be made by an independent panel, and they should be justified and based on the intended model use, not on convenience or likelihood of a successful validation outcome. Though it may be infeasible if the data are needed to build the model, studies chosen for validations should be independent. Models validated by partially dependent and even dependent validations can still be useful, however. Multiapplication models should be validated against as many ''landmark'' studies as possible (i.e., those used by experts as the basis for their understanding of the disease).

First, external validation usually requires multiple data sources, as a model will address various populations, interventions, and outcomes. Second, populations and care processes vary across settings, and it is important to explore how well a model simulates those. Third, it is important to validate the separate model parts. For example, a model can overestimate incidence, underestimate treatment effect, and still estimate mortality accurately, implying falsely that it is wholly valid.

Simulating the sources. The simulation should use information from the source, such as population characteristics, treatment protocols, and outcome definitions. Data on intermediate outcomes might be used if actual practice deviated from the intended design. For example, if a trial's design called for reducing LDL cholesterol to 100 mg/dL but it increased to 145 mg/dL, then it would be appropriate to use that information. The simulation setup should not be informed by the source health outcomes.

The simulation should be set up to match the source's circumstances as closely as possible, including setting, target populations, treatment and follow-up protocols, and outcome definitions. A mismatch in any aspect can affect interpretation of the validation. For example, if a data source reports ''myocardial infarctions,'' it is important to understand whether that includes only hospitalizations or also includes sudden deaths and silent infarctions. If the source includes all three but the model estimates only hospitalizations, then the event rates should not be expected to match. As a rule of thumb, if the investigators responsible for a data source thought it important to include an item, then modelers should try to

include the same level of detail in the simulation. To the extent possible, variables in the model that control behaviors should be set to match those in the source (e.g., treatment crossover). Modelers should identify aspects that cannot be matched and should discuss the implications for validation.

Parameters that define, for example, the condition's incidence and natural history, the effects of risk factors, the physiology, the occurrence of outcomes, and the effects of treatments should not be changed or ''refitted'' for each data source to achieve a good match. The model structure can be modified during building, but once a model is ready for external validation, it should not be modified further to fit a particular data source. Refitting could be an indication that a different model structure might be more appropriate.

Comparing outcomes. The comparison of simulation outcomes to the actual source events should include the same statistical methods used by source investigators. For example, if source results are presented as Kaplan-Meier curves, so too should the simulation's results. If the source measures outcomes at various follow-up times, the validation should include the same times.

Reporting should include a description of the data source, simulation setup (matching characteristics, follow-up, and any discrepancies), model and study uncertainty, and metrics used to assess consistency of model and study results.[56] Any factors known to affect outcomes (but not described in the source) or any that could not be simulated accurately should be reported, along with how any discrepancies might affect the validation results.

The next task is to explore quantitatively how uncertainty and discrepancies in actual versus simulated design affect the results and whether justifiable assumptions will diminish differences. If use of justifiable assumptions causes convergence, the model's results can be called ''consistent with'' actual results. The results of these sensitivity analyses should be reported along with the baseline analyses. The quantitative comparison of model results with actual ones will depend on the type of outcome. As it is rarely possible to match a source exactly, even highly accurate models might deviate, and results should be interpreted cautiously. For a summary measure, one can report the proportion of results within any specified bound. Information about the source and simulation sample size can be used to calculate whether a measure is statistically significant.

For multiapplication models modified over time or when new evidence becomes available, external validations need to be redone as an ongoing process.

Strengths and limitations. External validation tests the model's ability to calculate actual outcomes. This type of validation is used throughout health care (e.g., confidence in CT scans is based on comparisons of their results with actual physical findings) and, indeed, virtually every other scientific field.

External validation can address only parts covered by data sources. Even if a model accurately predicts a dozen clinical trials, there is no guarantee that it will be accurate for the next trial. Unless there happens to be a data source directly applicable to an analysis, external validation cannot directly validate it. This rarely occurs; if such a source existed, the model would not be needed. Another limitation is insufficient useful validation data. The number of data sources may be limited. Data sources may omit or be vague about some information needed to set up an external validation properly. Even when the information on the source's design exists, it may not accurately represent what happened, due to changes during the study. Even if protocols are described perfectly and followed rigorously, factors that vary across settings and affect outcomes may not be reported or may not even be known. Person-specific data may be unavailable, forcing use of aggregated data or assumed distributions of values. Accurate matching of aggregated results may not validate results for subpopulations.

Though informal sources are attractive because they represent ''real practice,'' their use is especially problematic because without a formal design, it is very difficult to determine what actually happened, given population turnover, practice pattern variations, selection biases, confounding, incomplete performance and adherence, and staggered adoption of new interventions. Many of these factors are not measured, and even when measured, they can be very difficult to simulate.

Another limitation is that the model might not include all elements needed to accurately simulate a source. It might not include all risk factors or comorbidities; all patient, physician, hospital, and health care system care processes or behaviors; or all features needed to calculate outcomes precisely as defined in a source protocol.

External validation is even more problematic for resource use and costs. Because of practice pattern variations, resource use triggered by clinical events differs across settings even when event rates are

similar. As unit costs can vary widely across settings, costs are subject to similar problems.

Best practices

VII-6 There should be a formal process for conducting external validation that includes the following:

- Systematic identification of
  - Suitable data sources
  - Justification of the selection
  - Specification of whether a data source is dependent, partially dependent, or independent
  - Description of which model parts are evaluated by each source
- Simulation of each source
- Comparison of results, including descriptions of
  - Data source
  - Simulation setup
  - Discrepancies between source and simulation and their implications
  - Discrepancies between simulation and observed results
  - Sensitivity analyses
- Quantitative measures of how well the model's results match the source outcomes

VII-7 Modelers should make available on request a description of the external validation process and results, identify model parts that cannot be validated given lack of suitable sources, and describe how uncertainty about those parts was addressed.

VII-8 For multiapplication models, modelers should describe criteria for determining when validations should be repeated and/or expanded.

Predictive Validation

Predictive validation involves identifying an opportunity in which a study design can be specified, simulating that design, recording the predicted outcomes, waiting for events to unfold, and comparing them with predictions.[13,57] This process is most easily envisioned for clinical trials that have published their designs but not yet reported results, but it can also be applied to cohort studies still in progress.

Strengths and limitations. Predictive validation is the most desirable type, as it corresponds best to modeling's purpose: predicting what will happen. It also ensures a completely independent validation, allowing no opportunities for altering the model to fit observed results. A limitation is that the results are necessarily in the future and rarely in time to be helpful for immediate decisions. They also require that there be a trial planned or in progress applicable to the decision at hand. Many models are built to synthesize the best available evidence and illuminate a policy decision for which no trial is ongoing, planned, or even feasible. At best, this validation method is applicable only for short-term outcomes when research is feasible.

This method is also subject to all the limitations of external validations—in particular, changes or breaches in design and factors outside the control of the original study design, such as the introduction of new technologies or changes in care practices. Thus, the best use of predictive validation is to simulate a clinical trial or other suitable data source that was initiated in the past whose results are not yet known but will be announced in the near future. This type of validation is most useful for multiple-use models that are expected to be in service after the source's results are revealed.

Best practices

VII-9 When feasible with respect to the decision being addressed and a future source's availability, a model should be tested for its prediction of future events. Builders of multiple-use models should seek opportunities to conduct predictive validations.

Interpretation of Validations

Ultimately, whether a model is sufficiently valid or accurate for a particular application must be determined by those who would use its results. The best practices described here are intended to provide users with the information needed to determine how useful a model and its results can be expected to be for their purposes. We recommend that users of a model examine validation results with four criteria:

- rigor of the process;
- quantity and quality of sources used (how well they represents the model's proposed use);
- model's ability to simulate sources in appropriate detail; and
- how closely results match observed outcomes, initially and after making justifiable assumptions about uncertain elements.

CONCLUSIONS

We have described methods and recommended best practices for making models transparent and

for validating them. These principles are essential for enabling readers and potential users to understand how a model works and to judge its expected accuracy when applied to particular problems. Not all models will be able to achieve all these best practices, and inability to do so does not necessarily imply that a model is not useful. Modelers should strive, however, to achieve these best practices. Beyond the limitations of transparency and validation described above, it is important to understand that models are only that; they are not reality. Models are developed to help decision makers when the questions are too complex for the human mind. Well-described and well-validated models can provide invaluable insights that cannot be obtained otherwise.

NOTES

1. the term ''reader'' describes anyone who needs to evaluate a model, including journal reviewers, journal readers, and users of a model's results
2. Peer reviewers should keep the technical documentation confidential as a matter of policy

REFERENCES

1. 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. Med Decis Making. 2012;32(5):667–677.

2. 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. Med Decis Making. 2012;32(5):678–689.

3. Siebert U, Alagoz O, Bayoumi AM, et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–3. Med Decis Making. 2012;32(5):690–700.

4. Karnon J, Stahl J, Alan B, Caro JJ, Mar J, Möller J. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–4. Med Decis Making. 2012; 32(5):701–711.

5. Pitman R, Fisman D, Zaric GS, et al. Dynamic transmission modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group–5. Med Decis Making. 2012;32(5): 712–721.

6. Briggs AH, Weinstein MC, Fenwick EAL, Karnon J, Sculpher MJ, Paltiel AD. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group–6. Med Decis Making. 2012; 32(5):722–732.

7. Goeree R, O'Brien BJ, Blackhouse G. Principles of good modeling practice in healthcare cost-effectiveness studies. Expert Rev Pharmacoecon Outcomes Res. 2004;4:189–98.

8. Karnon J, Goyder E, Tappenden P, et al. A review and critique of modelling in prioritising and designing screening programmes. Health Technol Assess. 2007;11:1–145.

9. Kopec JA, Fines P, Manuel DG, et al. Validation of population-based disease simulation models: a review of concepts and methods. BMC Public Health. 2010;10:710.

10. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. Pharmacoeconomics 2006;24:355–71.

11. 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;8:1–158.

12. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models: a suggested framework and example of application. Pharmacoeconomics. 2000;17: 461–77.

13. Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices—Modeling Studies. Value Health. 2003;6(1):9–17.

14. Weinstein MC, Toy EL, Sandberg EA, et al. Modeling for health care and other policy decisions: uses, roles, and validity. Value in Health 2001;4:348–61.

15. Weinstein MC, Coxson PG, Williams LW, et al. Forecasting coronary heart disease incidence, mortality, and cost: the coronary heart disease policy model. Am J Pub Health. 1987;77:1417–26.

16. Owens DK. Analytic tools for public health decision making. Med Decis Making. 2002;22:S3–10.

17. Freedberg KA, Losina E, Weinstein MC, et al. The cost effectiveness of combination antiretroviral therapy for HIV disease. N Engl J Med. 2001;344:824–31.

18. Goldie SJ, Yazdanpanah Y, Losina E, et al. Cost-effectiveness of HIV treatment in resource-poor settings: the case of Cote d'Ivoire. N Engl J Med. 2006;355:1141–53.

19. Paltiel AD, Weinstein MC, Kimmel AD, et al. Expanded screening for HIV in the United States: an analysis of cost-effectiveness. N Engl J Med. 2005;352:586–95.

20. Walensky RP, Wolf LL, Wood R, et al. When to start antiretroviral therapy in resource-limited settings. Ann Int Med. 2009;151:157–66.

21. Weinstein MC, Goldie SJ, Losina E, et al. Use of genotypic resistance testing to guide HIV therapy: clinical impact and cost-effectiveness. Ann Int Med. 2001;134:440–50.

22. Owens DK, Harris RA, Scott PM, et al. Screening surgeons for HIV infection: a cost-effectiveness analysis. Ann Int Med. 1995; 122:641–52.

23. Sanders GD, Hlatky MA, Owens DK. Cost-effectiveness of implantable cardioverter-defibrillators. N Engl J Med. 2005;353: 1471–80.

24. Schlessinger L, Eddy DM. Archimedes: a new model for simulating health care systems. The mathematical formulation. J Biomed Infor. 2002;35:37–50.

25. Eddy DM, Schlessinger L. Validation of the Archimedes diabetes model. Diabetes Care. 2003;26:3102–10.

26. Stern M, Williams K, Eddy DM, et al. Validation of prediction of diabetes by the Archimedes model and comparison with other predicting models. Diab Care. 2008;31:1670–1.

27. Kahn R, Alperin P, Eddy DM, et al. Age at initiation and frequency of screening to detect type 2 diabetes: a cost-effectiveness analysis [erratum in Lancet 2010;375:1346]. Lancet. 2010;375:1365–74.

28. Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. Ann Intern Med. 2007;146:450–3.

29. Morin A, Urban J, Adams PD, et al. Research priorities: shining light into black boxes. Science. 2012;336:159–60.

30. Peng RD. Reproducible research and biostatistics. Biostatistics. 2009;10:405–8.

31. Peng RD. Reproducible research in computational science. Science. 2011;334:1226–7.

32. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. Am J Epidemiol. 2006;163:783–9.

33. Institute of Medicine. Evolution of Translational Omics Lessons Learned and the Path Forward. Washington, DC: Institute of Medicine; 2012.

34. DeAngelis CD, Fontanarosa PB. The importance of independent academic statistical analysis. Biostatistics. 2010;11:383–420.

35. International Society for Pharmacoeconomics and Outcomes Research. Value in Health guide for authors. Accessed from: http://www.ispor.org/publications/value/submit.asp

36. 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.

37. BMJ. Health economics. Accessed from: http://www.bmj.com/about-bmj/resources-authors/article-types/research/health-economics

38. Groves T. The wider concept of data sharing: view from the BMJ. Biostatistics. 2010;11:391–2.

39. Baggerly K. Disclose all data in publications. Nature. 2010;467:401.

40. Keiding N. Reproducible research and the substantive context. Biostatistics. 2010;11:376–8.

41. JAMA. Instructions for authors. Accessed from: http://jama.ama-assn.org/site/misc/ifora.xhtml

42. McHaney R. Computer Simulation: A Practical Perspective. San Diego, CA: Academic Press; 1991.

43. Barter PJ, Caulfield M, Eriksson M, et al. Effects of torcetrapib in patients at high risk for coronary events. N Eng J Med. 2007;357:2109–22.

44. Bratley P, Fox BL, Schrage LE. A Guide to Simulation. 2nd ed. New York: Springer-Veriag; 1987.

45. Law AM, Kelton WD. Simulation Modeling and Analysis. 4th ed. New York: McGraw-Hill; 2007.

46. Thesen A, Travis LE. Simulation for Decision Making. St. Paul, MN: West Publishing; 1992.

47. Drummond MF, Barbieri M, Wong JB. Analytic choices in economic models of treatments for rheumatoid arthritis: what makes a difference? Med Decis Making. 2005;25:520–33.

48. Berry DA, Cronin KA, Plevritis SK, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. New Eng J Med. 2005;353:1784–92.

49. Mount Hood 4 Modeling G. Computer modeling of diabetes and its complications: a report on the Fourth Mount Hood Challenge Meeting. Diabetes Care. 2007;30:1638–46.

50. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, et al. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Int Med. 2008;149:659–69.

51. Stout NK, Knudsen AB, Kong CY, et al. Calibration methods used in cancer simulation models and suggested reporting guidelines. Pharmacoeconomics. 2009;27:533–45.

52. Royle P, Waugh N. Literature searching for clinical and cost-effectiveness studies used in health technology assessment reports carried out for the National Institute for Clinical Excellence appraisal system. Health Technol Assess. 2003;7:1–51.

53. Haynes RB, McKibbon KA, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ. 2005;330:1179.

54. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ. 2004;328:1040.

55. Montori VM, Wilczynski NL, Morgan D, et al. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. BMJ. 2005;330:68.

56. Goldhaber-Fiebert JD, Stout NK, Goldie SJ. Empirically evaluating decision-analytic models. Value Health. 2010;13:667–74.

57. Eddy DM. The frequency of cervical cancer screening: comparison of a mathematical model with empirical data. Cancer. 1987;60:1117–22.