Introduction

The purpose of healthcare models is to provide decision makers with quantitative information about the consequences of the options they are considering. In order for a model to be useful for this purpose, decision makers need to know how much confidence they can place in the model’s results. More specifically, they need to know how accurately the model’s results can be expected to predict the outcomes in which they are interested. They can then take that information into account when deciding how to use the results of the model.

There are three main ways by which decision makers can gain such confidence. One is for them to understand the structure and assumptions of the model, both in general and as applied to a specific problem. Subject to their expertise, their ability to do this is determined by the transparency of the model—the extent to which the structure and assumptions are described. The second is through validation of the model—subjecting it to tests such as comparisons of the model’s results with events observed in reality. The third is through the report of an analysis—information about the problem being addressed, the structure and assumptions of the model, the results and sensitivity analyses, limitations of the analysis, and conclusions. All three of these are necessary. Other papers in this series describe best practices for reporting the methods and results of analyses conducted using particular types of models. This paper addresses best practices for transparency and validations, and is intended to apply to all types of models.

Some healthcare models are intended to be “general” or “multi-application” in the sense that with appropriate modifications they can be used to address a range of problems. For example, an “HIV model” could be used repeatedly to address different questions relating to that condition. Other models are built for single specific applications and are not intended to be reused. The methods described in this paper are applicable to both types of models. For a multi-application model, transparency, validation, and reporting are ongoing processes. The multi-application model is described (transparency) and validated, and the descriptions and validations are continuously updated as the model and science evolve. In addition, each application of the model is described, validated, and reported as it is done. For a single-application model, the description and validation of the model, 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 they need in order to determine how much confidence they should place in, and how they should use, the model’s results. Every model today should be able to achieve the best practices we recommend for transparency. However, we recognize that not all models today will be able to achieve all the best practices we recommend for validation. Rather than water down our recommendations, we have described the level that top-quality models can achieve and that all models should strive to achieve. If a model does not meet all the recommended best validation practices, its developers should describe their process for conducting validations and the level of validation their model is able to achieve.

Transparency

In this context, transparency refers to the extent to which interested parties can review the structure, equations, parameter values, and assumptions of a model. It does not refer to the formulation, conduct, or results of a particular analysis; those issues are discussed in the sections on reporting in other papers in this series. Transparency serves two purposes. One is to provide a non-quantitative description of the model to readers who want to understand in a general way how a model works. Another is to provide technical information to readers who want to evaluate a model at higher level of mathematical and programming detail. (We will use the term “reader” to describe anyone who needs to evaluate a model, including journal reviewers, journal readers, and potential users of a model’s results.) 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 of detail appropriate to their expertise and needs.

Non-technical documentation. Consistent with the dual purpose of transparency, it is useful to distinguish between “non-technical” and “technical” documentation. Non-technical documentation is written for readers who are familiar with the clinical and policy–related elements of the problem area, but do not have sufficient skills to review technical aspects of the model such as equations, parameter estimation, or programming methods. The non-technical documentation should be accessible to any interested reader and should include the following information about the model and its inputs and outputs:

- general description of the model and its purpose
- description of the types of applications the model is designed to address (e.g., forecasting of short-term costs, classical cost-effectiveness analysis)
- sources of funding and their role, if any, beyond providing financial support
- graphical representation (e.g., influence diagram) or other description of the structure of the model and the components that define it and determine its performance
- list of the inputs, outputs, other parameters, and equations
- descriptions and citations of the sources used to write the equations
- description of how the data sources were identified and selected
- examples of equations (optional)
- description of the methods used to validate the model, customize it to specific applications and settings, and explore the effects of uncertainty
- summary of the results of validations
- description of the main limitations of the model for its intended applications
- reference to a technical description of the model, subject to conditions that will be described below

The non-technical documentation will give readers an overview of the model and what it does, but except for the simplest models will not contain sufficient information to enable readers to determine precisely how it does it or to replicate it. To determine how a model works, technical documentation is needed.

Technical documentation. The most direct and widely proposed method for achieving full technical transparency is to enable readers to see documents that detail the model, including its structure, components, equations, and code. The documentation should be sufficiently detailed to enable readers who have the necessary expertise and resources to reproduce the model. The benefits of full technical documentation are obvious; it is the only way readers can understand how the model works.

However provision of technical documentation is subject to some conditions and limitations. First, access to technical documentation should be provided in a way that enables protection of intellectual property. Building a model can require
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. Second, because few journals are willing to publish mathematical details of models, even as appendices, most technical documents will need to be made accessible by other means. Third, most multi-application models change over time; they are expanded and updated to incorporate new information and advances in healthcare technologies. Therefore, technical documents will need to be updated periodically. Fourth, the equations and detailed structure of a model will mean little to readers who do not have the necessary technical background. Even for readers who have the necessary background, reviewing a model can take considerable time. Furthermore, it is very difficult to understand how accurate a model is simply by 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 the code, which for large models or models that require advanced computing methods (e.g. distributed computing), can be very difficult. 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 of the model 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 of the model applicable to a particular analysis. Having said this, it is important to note that providing such access can be very expensive, not just due to the cost to build the copy and interfaces, but to providing support to ensure that the model is used and interpreted accurately. Modelers who make working copies of their models 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 vs. confidential documentation.** To address the conflicting needs of transparency versus feasibility and intellectual property there needs to be a distinction between information put in the public domain without restriction (“public documentation”) versus information made available to readers under agreements that protect intellectual property (“confidential documentation”). For public documentation, modelers should provide the non-technical description of the model, as outlined above, to anyone who asks. In addition, at their discretion modelers can put in the public domain any technical documentation they choose, or a working copy of the model. Regarding confidential documentation, modelers should be prepared to provide full technical documentation (along with access to a working copy, if available) to readers designated either by a journal to which the modelers are submitting a paper or by an organization to which the modelers are providing the model for decision making, under agreements that protect intellectual property. Peer reviewers for grant applications or publications should keep the technical documentation confidential as a matter of policy. Providing readers with technical information under conditions of confidentiality is consistent with JAMA’s requirements for the review of models.1 Because of the size and technical nature of the documentation, because of 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 technical documentation of a model be included in the published report of an analysis

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

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. The need for advanced mathematical training by itself prohibits this. Those who do not have the appropriate training or time to study a model must understand that a model is not necessarily flawed if it is not transparent to them. It is also important to stress that except for the simplest of models, transparency by itself does not document a model’s accuracy for any particular problem. A model can be totally transparent but give the wrong answer. The erroneous formula Distance = Rate/Time is an example. Conversely, a model can be quite non-transparent to most readers, but correct. The Schrödinger equation is an example: $i\hbar \frac{\partial}{\partial t} \Psi = \hat{H} \Psi$. Ultimately, what matters is whether or not a model accurately
predicts what occurs in reality. Thus transparency and validation are inextricably linked and ultimately it is the validation that counts most in terms of helping readers gain confidence in the accuracy of a model. An analogy is that the equations used to convert CT scans into images are not transparent to any physician, yet physicians use CT scans all the time. They are willing to do so because when they operate based on a CT scan that shows a lump, they almost always find a lump. The key to developing confidence in a model is not just to study its equations and code, but to examine whether or not it accurately calculates the outcomes of interest. This is the role of validation.

**Validation**

**Importance of validation.** Models are developed to help answer questions by predicting outcomes of interest, in populations and settings of interest, with interventions of interest. Validation is a set of methods for judging a model’s accuracy in making such predictions. For multi-application models it is also a process for continuously improving the model’s accuracy. That information can then be used by decision makers to determine the applicability of the model’s results 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. It is the only way to measure the actual accuracy of a model’s results.

If a model is intended to be used for multiple applications, a distinction should be made between validating the model in a general sense (e.g., as a "diabetes model" or "heart disease model") versus validating the model for a specific application (e.g., the effect of a particular weight loss drug on glycemia levels in people with diabetes). If a model is built for only one application (a single-application model), validation can be limited to that particular application. Validation is vital for both multi- and single-application models.

It is not possible to specify precise criteria a model must meet in order to be declared "valid" or “a validated model,” as though validity were a property of a model that applies to all of its applications and uses for all time. One reason is that applications vary and a model can have different levels of validity for different applications. Thus, the concept of validation should apply to particular applications, not to the model itself. A second reason is that the required degree of accuracy depends on the nature of the question being addressed. For example, much less accuracy is needed to answer the question “Will this intervention increase or decrease costs?” than to answer the question “How much will this intervention cost?” or “How many years will it take before this intervention breaks even?” A third reason applies to multi-application models – they can and should change over time to keep up to date with new medical science, technologies, and evidence.

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 the results of a model vary depending on uncertainty about a model’s components one-by-one and in combination, as discussed elsewhere in this series. However sensitivity analysis by itself does not evaluate how accurate a model is in simulating what actually occurs in reality. Sensitivity analysis is an important component of 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, its assumptions, and its applications correspond to current science and evidence, as judged by people who have expertise in the problem areas to which the model is being applied. Verification or internal validity addresses whether the parts of a model behave as intended and the model has been implemented (e.g., coded) correctly. Cross validation involves comparing a model with other models and determining the extent to which different models calculate similar results. In external validation a model is used to simulate events that have occurred in the real world, such as a clinical trial, and compare the model’s predicted outcomes with the real-world outcomes. Predictive validity involves using a model to predict events that have not yet occurred and, after the passage of time, comparing the predicted outcomes to the actual outcomes. Each type of validation has methods, strengths, limitations, and best practices.

**Face validity.** Four aspects of a model and its applications are particularly important for face validity: the structure of the model, data sources, the way the model is set up for a particular application (problem formulation), and the results of an application.

Methods. Face validity is a subjective judgment; people who have expertise in the problem area study the problem formulation, model structure, sources of evidence, assumptions, and results, and evaluate how well each of those aspects
of the model corresponds to their understanding of the pertinent medical science, the best available evidence, and the
clinical or administrative question that is being addressed. Information about the model and supporting evidence is
obtained from the documents provided by the modelers (discussed above). Information about the problem formulation and
results is obtained from the report of the application.

Specific questions depend on which aspect of the model or its application is being evaluated. For the structure, important
questions are whether the model includes all aspects of reality the expert considers important, and whether they are
related in ways that are consistent with medical science. For evidence, those judging face validity ask whether the best
available sources were used. For problem formulation, pertinent questions include whether the setting, population,
interventions, outcomes, assumptions, and time horizons simulated by the model correspond to those of interest. For
results, the main questions are whether they match experts’ expectations and, if not, whether the model can plausibly
explain them. If there are perceived weaknesses in any of these aspects of a model, the assessment should examine how
well the authors have reported the discrepancies, explained them, and described their potential effects 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, people in the same organization (e.g., university, agency, or company) who did not build the model, and/or
consultants chosen from outside the organization. Any reader can perform his or her own evaluation of face validity. Any
reader or organization can identify other experts and have them perform an evaluation. Peer review typically includes an
evaluation of face validity. Because face validity is a subjective judgment, the evaluators should have no stake in the
problem the model is being used to analyze. If the evaluators do have a stake in the results of an analysis, the structure,
evidence, and problem formulation should be assessed without knowing the results. In our judgment, as presently
practiced peer review of a model for publication is not sufficiently consistent to be relied on for determining face validity.

Strengths and limitations. The main strengths of face validation are that it helps ensure that a model is constructed and
used in a way that corresponds with the most current medical science and best available evidence. This process enhances
credibility with experts and increases acceptance of the model’s results. Additional benefits are that the act of evaluating a
model for face validity can raise questions and force thinking that improves the model. If a model’s results are
counterintuitive but justified, exploring the causes of the results can identify new hypotheses and stimulate additional data
collection and research.

However, face validation has several limitations. All models simplify reality, many to a very large extent. Thus, the
structures of many models are not completely consistent with current medical knowledge 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 occurs in state-
transition models. Physicians know that age, sex, and smoking are not “added” in any real sense, although they are
weighted and added in regression equations used in many models. Despite these simplifications, for properly selected
problems state-transition models and regression equations can be sufficiently accurate to meet the needs of decision-
makers. It is very difficult if not impossible for anyone to determine in their heads (subjectively) whether a model has been
properly simplified, oversimplified, or under-simplified for a particular problem.

Another limitation is that current medical evidence is incomplete and 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 current
misconceptions into a model. An example is that, until recently virtually all experts believed that raising HDL cholesterol
levels reduced cardiovascular disease events; a model that included that effect would have had high face validity, and a
model that included the opposite effect would have had low face validity. Yet a recent clinical trial contradicted those
beliefs.4

A third limitation is that there are no clear lines or unambiguous criteria to apply to judgments about a model or its
application. Lacking these, it is easy for anyone who has a stake in the model or result of an analysis to be swayed. More
specifically, virtually every modeler would say that his or her model has face validity. And anyone with a stake in the result
of an analysis will have a strong bias to accept a model if he or she likes its results, and otherwise reject it or simply ignore
it.
Best practices: Face Validity. Validation of a model should include an evaluation of face validity of the structure, evidence, problem formulation, and results of the model. A description of the process used to evaluate face validity should be made available on request. Evaluation of face validity should be made by people who have expertise in the problem area, but are impartial to the results of an analysis. If face validation raises questions about a model, these issues should be discussed by the modelers in their report of an analysis.

**Verification.** This type of validity, which is also called internal validity, internal consistency, or technical validity⁶ ⁷ ⁸, examines the extent to which the mathematical calculations are performed correctly and are consistent with the specifications of the model.

**Methods.** The methods will depend on the complexity of the model. In general, verification involves two main steps: validating the individual equations, and validating that the code accurately implements the equations. Individual equations and parameters in a model should be validated against the sources used to build the model and against independent sources if available. Examples are: curve fitting for equations; simulation of data sources that apply to parts of a model (e.g., use of epidemiological data to validate equations for incidence of a disease⁹); The accuracy of coding should be checked using state-of-the-art quality assurance and quality control methods for software engineering. Examples of techniques are: 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; verification of 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; sensitivity analysis; extreme value analysis¹⁰; trace analysis, in which individual events and their timing are tracked; and identification of unnecessary detail that might increase the likelihood of errors¹¹. The choice of methods should be appropriate for the complexity of a model.

**Strengths and limitations.** The main strength of this type of validation is that it helps ensure that there are no errors or “bugs” in the calculations. The main limitation is that it does not evaluate the accuracy of the structure of the model. 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 quite inaccurate in estimating the outcomes of interest if the structure of the model itself is poorly chosen. For example, if a question involving distance (D), rate (R) and time (T) is set up as D = α + β₁R₁ + β₂ T, instead of D=R*T, the parameters α, β₁, and β₂ can be estimated using state-of-the-art statistical methods and the equation can be coded correctly, but the results can be wrong. Verification of code will not identify such problems.

**Best practices: Verification.** Models should be subjected to rigorous verification. The verification methods should be described in the non-technical documentation of the model. The pertinent results of verification should be made available on request.

**Cross validation.** This method, which has also been called external consistency and external convergence testing, involves giving the same problem to different models and comparing their results. For example, six different cardiovascular disease models might be used to calculate the 10-year rate of non-fatal myocardial infarctions under a standardized set of conditions such as the same population, interventions, and levels of performance and adherence. The extent of differences among the models’ results and sources of the differences are then examined.

**Strengths and limitations.** Confidence in a result is increased if similar results are calculated by different models using different methods. An example would be the dating of an archeological find by both carbon 14 and tree ring methods. Comparisons across models can also be useful for methodological purposes.

However, cross validations have several limitations. First, this type of validation is only possible when there are multiple models capable of analyzing the same problem. Second, the modelers have to agree to participate. Third, agreement of results does not necessarily mean that the models are valid. The meaningfulness of this type of validation depends on the degree to which the methods and data sources of the different models are independent, as in the archeology example. In healthcare modeling there is often a high degree of dependency among models in that they may use similar frameworks and the same data sources. Some models even draw on parameters used in other models published earlier. Similar results could as easily mean that all the models are inaccurate as that they are accurate. Fourth, when results are different, there is often no way to determine which, if any model is correct.
Best practices: Cross validations. Modelers should search for previously published modeling analyses of the same or similar problems and discuss insights gained from similarities and differences in results.

External validation. This type of validation compares the results of a model to data on actual events\textsuperscript{12 13}. It is conducted by using the model to simulate a set of events that have occurred, such as those documented in a clinical trial, and examining how well the results correspond. For multi-application models, external validations can be applied to the model in a general sense (without any particular application in mind) as well as to each of its applications. For single-application models, external validation should be applied to the particular application. External validation can also be applied to parts of a model such as the methods for creating simulated populations, 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. For reasons that will be discussed in the section on limitations, external validations are not applicable to utilization and procedure outcomes (e.g. admissions, bypass rates) or financial costs.

Because the fundamental purpose of a model is to help decision-makers anticipate what will occur if they take certain actions, this and predictive validation are the types of validation that most closely correspond to that purpose and are therefore the most important. For these reasons, we describe the methods, strengths, limitations, and best practices for external validation in greater detail than for other types.

Methods. The principle underlying external validation is that a model’s calculations should be compared with what has been observed in settings and other circumstances that correspond as closely as possible to the proposed uses of the model. Every part of a model that affects its results, as well as the model as a whole, should be validated against as many actual experiences as possible. Examples of validation of parts are that an epidemiological study can be used to validate a model’s incidence equations, and the progression of biomarkers in a trial’s control group can be used to validate its physiologic equations. In contrast, simulation of an entire clinical trial will test several or all parts of a model at once, and simulation of multiple trials will test the model’s accuracy in calculating multiple outcomes, in multiple populations, treated with multiple interventions.

External validations involve three main steps: identifying the data sources the model should simulate, conducting the simulation, and comparing results.

To identify data sources, there are two main requirements: the source must contain data that are applicable to the model (or its parts) and its intended uses, and the data source must be sufficiently well described to enable replicating the design and progression, and calculating the outcomes of interest. The “design” of a data source includes information about the setting, population, treatment protocols, follow-up protocols, and outcomes. By “progression”, we mean any changes in the design or conduct of the data source (e.g. a clinical trial) over the follow-up period. Examples of data sources are population statistics, epidemiologic studies, clinical trials, claims data, and electronic health records.

Sources of data for validations can be either formal or informal. In this context a “formal” data source is one intended for research purposes and includes explicit planning and description of such things as: 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 (e.g., Kaplan-Meier curves). We will use the term “informal” to describe a source that is intended primarily for other purposes and does not include explicit descriptions of all the elements of formal sources. Examples of the latter are clinical records and claims data. The distinction is important because a central element of external validation is simulation of the data source, which can be difficult without explicit planning and descriptions.

Depending on the extent to which data from a particular source were used to help build a model, a validation using that source data can be dependent, partially-dependent, or independent. A validation is considered dependent if the data source was used to estimate parameters for the model relating to the same outcomes used to validate the model. A validation is considered partially dependent if the data 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 data source can be dependent for one outcome but independent for another. A validation is considered independent if no information from the source, beyond the information required to set up the simulation, was used to build any part of the model. Independent validations can be blinded or unblinded. A validation is considered blinded if those performing the validation had no information about the
outcomes in the data source that the model is calculating. However, even when a data source is unblinded, those conducting an independent validation should not allow information about the outcomes to influence the validation.

Data sources should be identified by conducting a formal search of pertinent databases using established methods\textsuperscript{14}, identifying sources that involve settings, populations, interventions, and outcomes similar to those the model is intended to analyze. From among these, sources with the best designs (e.g., large size, representative population, formal protocols, detailed reporting, and recent date) should be selected. For multi-application models there are often “landmark” surveys and trials that experts already use as the basis for their own understanding of the disease. Multi-application models should be validated against as many of these as possible. For single-application models, sources particularly pertinent to that application should be chosen, in addition to any landmark trials that are closely related to the application.

To the greatest extent possible, studies chosen for validations should be independent, although that may not be feasible, as there may not be sufficient data to both build a model and conduct independent validations that cover all of the important parts of a model and its applications. Models validated by partially-dependent and even dependent validations can still be very useful, and the lack of independent validation does not necessarily mean that a model has no value. For example, a model that successfully simulates the landmark trials in a field can have great value, even if modelers used some information from those trials to build it. An analogy is that the knowledge of experts is based on existing evidence and therefore highly dependent, and yet undeniably useful.

Ideally, the validation plan and sources should be chosen before the results of an analysis are known, or possibly even before the model is built (or modified, in the case of a multi-application model). The choices of data sources should be based on the intended use of the model, not on convenience or likelihood of a successful validation outcome. The validation sources should have designs that match as closely as possible the questions and outcomes that the model is intended to address. Ideally the data sources will be chosen by an independent panel, which will also monitor the validation process and review the results.

External validation of a model almost always requires multiple data sources, for three main reasons. One is that unless a model is extremely simple, it will be designed to address a variety of populations, interventions and outcomes. It is important to validate as many as possible. Second, populations and care processes vary in different 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 and underestimate treatment effect, and still end up estimating mortality accurately, giving the false impression that it is accurate. A model part is considered to be tested if its equations are used, possibly in combination with the equations of other parts, to calculate the outcomes of interest. Thus, external validation should include data sources that address each part of a model and combinations of parts, as well as the whole model. For the overall model, it is important to perform multiple validations that cross the intended applications in the sense of involving a range of populations, interventions, outcomes, and time horizons. The number of data sources required to validate a model will depend on the quality and quantity of the available sources, the model, and the proposed applications, but as many as possible should be used. The number and choices should be explained. To the greatest extent possible those selected should not be dependent (they should either be independent or partially independent).

The second step is to use the model to simulate the data sources. The simulation is set up using information from the data source such as characteristics of the population, treatment protocols, and definitions of outcomes. Data on intermediate outcomes might be used if they indicate that actual practice deviated from the intended design. For example, if a trial’s design called for reducing LDL cholesterol levels to 100 mg/dl but the levels gradually increased to 145 mg/dl, then it would be appropriate to use that information. The setup of the simulation should not be informed by any information about the outcomes of the data source that the model is trying to calculate.

The simulation should be set up to match the circumstances of the data source as closely as possible, including the setting, target populations, treatment protocols, 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 if that includes only hospitalizations or also includes sudden deaths and silent myocardial infarctions. If the validation source includes all three but the model estimates only numbers of hospitalizations, then the overall event rates should not be expected to match. As a rule of thumb, if the investigators responsible for a data source thought it was important to include an item in a trial’s protocol, 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 the behaviors described in the data source (e.g. cross-over of treatment groups).

It is unlikely that a model will be able to match every aspect of a data source. Modelers should identify aspects that cannot be matched and discuss the implications for validation. Factors known to affect the outcomes of interest, but not reported in the data source, should be explored through sensitivity analysis. For example, a clinical trial of a cholesterol-lowering treatment may not report the proportion of patients on aspirin, which can affect rates of coronary events\textsuperscript{15}. For validation, aspirin use could be based on other sources, and uncertainty about aspirin use examined across a range.

When validating against multiple data sources, it is important that only the parameters relating to the source’s design and progression be changed. Parameters that define such things as the condition’s incidence and natural history, effects of risk factors, physiology, occurrence of outcomes, and effects of treatments should not be changed or "refitted" for each data source in order 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. If such refitting is necessary, there is no way to determine the parameters for a new application. Refitting is also an indication that a different model structure might be more appropriate.

After the simulation of a data source has been completed, the third step is to compare the outcomes of the simulation to the events that occurred in reality. The comparison should include the same statistical methods used by investigators of the study to present the results. For example, if a trial’s results are presented as Kaplan Meier curves, the simulation’s results should be presented in the same way. If the trial defines "major coronary events" to include acute coronary syndrome, then the model should include acute coronary syndrome in its calculation of that outcome. If the data source contains information about outcomes measured at different follow-up times, the validation should include all the reported times. In addition to reporting the outcomes as they are reported in the data source, those validating a model may choose to report additional outcomes that they can calculate from the data source.

For each data source, modelers should compare the model’s results to the actual results. This has two parts. The first is a qualitative description of the data source, setup of the simulation, and any discrepancies between the two. This should include descriptions of the data source’s setting, population, interventions (including care processes and behaviors), follow-up protocols, definitions of outcomes, and results; how each of these aspects of the data source was represented in the setup of the simulation; any factors that are known to affect the outcomes observed in the data source, but that are either not described in the design or follow-up, or could not be simulated accurately by the model; how each of the elements of the design was addressed in the simulation; and how any discrepancies between the model’s simulation and the real trial might affect the comparison of results of the reference case to the results of the data source.

The second part is a quantitative comparison of the model’s (reference case) results and the actual results, and a sensitivity analysis to determine of how reasonable assumptions about the discrepancies affect the comparison. The method used to make the comparison will depend on the type of outcome being compared. For example, a measure of how closely a model’s results match the data source might be the ratio of calculated to observed hazards at each reported time; 1.0 representing a perfect match. The ratios should be interpreted with care. Depending on the extent to which there are discrepancies between the design and progress of the data source and the setup of the simulation, a mismatch might not indicate a flaw in the model. Conversely a good match might indicate a flaw if the setup is known to be an inaccurate representation of the data source. As it is rarely possible to match a data source exactly, even for highly accurate models, ratios will deviate by various amounts: ±0.1, ±0.2, or even more. For a summary measure one can report the proportion of ratios within any specified bound. Information about the sample sizes of the trial and simulation can be used to calculate whether a ratio is statistically significantly different from 1, or any set of bounds around 1.

These steps apply to a baseline case, where uncertain parameters are set to particular values (e.g., aspirin use is assumed to be zero, or silent myocardial infarctions are ignored). The next task is to explore quantitatively how uncertainty and discrepancies in actual vs. simulated design might affect the results, and whether justifiable assumptions will cause ratios of hazard rates to approach 1. As examples, aspirin use would be set to rates reported in some other data source from a similar setting and time period, or silent myocardial infarctions would be assumed to comprise about 35% of all myocardial infarctions in diabetics\textsuperscript{16}. If use of justifiable assumptions causes the ratio to converge on 1, the model’s results can be said to be “consistent with” actual results. The results of sensitivity analyses of this type and the ratios for a justifiable model should be reported along with the results of the baseline validation.
A final point is that for multi-application models that are modified over time, external validations need to be repeated and expanded whenever the model is changed or new evidence becomes available. For such models, external validation is an ongoing process, not a one-time event.

**Strengths.** The main strength of external validation is that it tests the ability of the model to calculate actual outcomes. It is also notable that this is the type of validation used throughout other aspects of healthcare (e.g., confidence in a CT scan), and indeed virtually every other scientific field.

**Limitations.** External validation can only address the parts of a model covered by the data sources. Even if a model accurately predicts a dozen clinical trial results, there is no guarantee that it will be accurate for the next trial. Healthcare is full of unknowns that cannot be predicted (modeled) from existing information. A related limitation is that external validations of existing data sources do not directly validate the results of a particular analysis, unless there happens to be a data source directly applicable to that analysis. This is very rare; if such a source existed, the model would not be needed.

Another potential limitation is insufficient useful validation data. The number of data sources may be small. Data sources may omit some pieces of information needed to set up an external validation properly. Even when the information on the design of a data source exists, it may not accurately represent what happened because of changes during the study, or vague descriptions. Even if protocols are described perfectly and followed rigorously, there are factors that vary across settings and affect outcomes but that are not reported, and may not even be known. Person-specific data may not be available from a data source, forcing the use of aggregated data or forcing assumptions about the distributions of values. Accurate matching of aggregated results will not validate results for subpopulations.

Use of informal sources such as clinical records and claims data is especially problematic. Those sources are attractive because they represent what happens in “real practice”, but without a formal design it can be very difficult to determine what actually happened because of such things as turnover of the population, variations in practice patterns, selection biases affecting which patients receive which treatments, confounding, incomplete performance and adherence, and staggered adoption of new interventions. Many of these factors are not measured. Even when measured, such factors can be very difficult to simulate.

Another set of limitations relates to the models; they might not include all the elements needed to accurately simulate a data source. Examples are that a model might not include all the important risk factors or co-morbidities that define a population, all the care processes or behaviors needed to simulate a treatment protocol or incomplete adherence, or all the outcomes needed to calculate outcomes precisely as defined in a trial’s protocol.

External validation is even more problematic for modeling resource use and costs. Because of variations in practice patterns, resource use triggered by clinical events differs across settings even if the rates of the clinical events are similar. Because unit costs can vary widely across settings, costs are subject to similar problems. Calculation of procedures and costs should be customized to particular settings.

**Best practices.** Builders of models should have a formal process for conducting external validation that includes:

- Systematic identification of suitable data sources; justification of the selection; specification of whether a data source is dependent, partially dependent, or independent; and description of which parts of the model are evaluated by each data source.
- Simulation of each data source and comparison of results.
- Comparison of results should include descriptions of:
  - Data source
  - Set up of the simulation
  - Discrepancies between the data source and simulation setup, and implications of the discrepancies
  - Comparisons of simulation results with observed results
  - Discussion of discrepancies between simulation results and observed results
  - Sensitivity analyses
- Quantitative measures of how well the model’s results match the outcomes observed in the data source

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

For multi-application models, modelers should describe criteria for determining when validations should be repeated and/or expanded.

**Predictive validity.** This type of validation is intended to determine how accurately a model can predict future events.

**Methods.** 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 events predicted. This process is most easily envisioned for clinical trials that have published their designs but not yet reported results.

Strengths and limitations. Predictive validation is the most desirable type of validation as it corresponds most closely to the purpose of modeling: 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 of predictive validations 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 that is applicable to the decision at hand. Many models are built to synthesize the best evidence available at any particular time, 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. A second problem is that this method is subject to all the limitations described for 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. Because of these limitations, 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 can also be useful for multiple-use models that are expected to be in service after the results of the data source are revealed.

**Best practices.** When feasible with respect to the decision being addressed and the availability of a future data source, a model should be tested for its prediction of future events. Builders of multiple-use models should seek opportunities to conduct predictive validations as part of their overall validation process.

**Interpretation of external validations.** Ultimately, whether a model is sufficiently valid or accurate for a particular application or decision must be determined by those who would use the model’s results. The best practices described herein are intended to provide potential users with the information needed to determine how useful a model and its results can be expected to be for their own intended purposes.

In addition to providing information about the accuracy of the model for matching events observed in reality, external validations also provide information about the ability of a model to represent aspects of reality that are important to whatever questions a potential user might have, at a level of detail the potential user considers important. To this end, we recommend that potential users of a model should examine the results of external validations with four main things in mind: rigor of the validation process; quantity and quality of data sources used to validate, where “quality” refers to how well the data sources represent the proposed uses of the model; ability of the model to accurately simulate the data source at a level of detail corresponding to the data source and proposed uses of the model; and how closely the model’s results match the 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 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 it is applied to particular problems. Not all models will be able to achieve all these best practices, and inability to achieve all of these practices does not necessarily mean that a model will not be useful. However modelers should strive to achieve these best practices.

Beyond the limitations of transparency and validation described above, it is important to understand that models are only models; they are not reality. Models are developed to help decision makers when the questions are too complex for the
unaided human mind. Well described and validated models can provide invaluable insights that cannot be obtained otherwise.

1 “Authors of reports of cost-effectiveness analyses and decision analyses must submit a copy of the decision tree comprising their model. This is for editorial evaluation and review, not necessarily for publication, unless it is included in the body of the manuscript.” http://jama.ama-assn.org/site/misc/ifora.xhtml accessed March 23, 2011.
10 (see for example http://onlinelibrary.wiley.com/doi/10.1046/j.1524-4733.2003.65241.x/full