Modeling for Health Care and Other Policy Decisions: Uses, Roles, and Validity

Milton C. Weinstein, PhD, Edmond L. Toy, ScM, Eileen A. Sandberg, ScM, Peter J. Neumann, ScD, John S. Evans, ScD, Karen M. Kuntz, ScD, John D. Graham, PhD, James K. Hammitt, PhD
Program on the Economic Evaluation of Medical Technology, Harvard Center for Risk Analysis, Harvard School of Public Health, Boston, MA

ABSTRACT

The role of models to support recommendations on the cost-effective use of medical technologies and pharmaceuticals is controversial. At the heart of the controversy is the degree to which experimental or other empirical evidence should be required prior to model use. The controversy stems in part from a misconception that the role of models is to establish truth rather than to guide clinical and policy decisions. In other domains of public policy that involve human life and health, such as environmental protection and defense strategy, models are generally accepted as decision aids, and many models have been formally incorporated into regulatory processes and governmental decision making. We formulate an analytical framework for evaluating the role of models as aids to decision making. Implications for the implementation of Section 114 of the Food and Drug Administration Modernization Act (FDAMA) are derived from this framework.

Keywords: cost-effectiveness analysis, models, pharmacoeconomics, validation.

Introduction

The use of models in health-care decision making is controversial. Resistance to models can come from physicians who claim that clinical judgment cannot be quantified, from empiricists who warn that input data can be inaccurate, from epidemiologists who worry that logical assumptions about cause and effect may be wrong, from technophobes who worry about hidden bugs lurking in black boxes, and from cynics who fear that proponents of a medical practice or product can manipulate models in hidden ways to mislead decision makers. There is an element of truth in all these concerns. Perhaps the concern is that models, by virtue of their quantitative nature and precise-looking computer outputs, may carry more influence than they should. For all these reasons, and because the stakes of health-care decisions are so great, models must be used in accordance with reasonable public policies concerning their use. We will argue that, although empirical tests of model predictions against retrospective and prospective data should be a part of such policies, they should not be erected as rigid and arbitrary barriers against the responsible use of models.

In the United States, the Food and Drug Administration (FDA) regulates information disseminated by drug companies regarding their products. The implementation of Section 114 of the FDA Modernization Act of 1997 (FDAMA) has spurred renewed interest in creating a balance between the flow of cost-effectiveness information to decision makers and protection against unwarranted claims. Other countries, in Europe and elsewhere, are also assessing the role of models as evidence for cost-effectiveness claims. The fundamental policy question facing these agencies is under which conditions drug manufacturers should be permitted to use models as the basis for claims of cost-effectiveness.

The controversy over validation of models originates in part from different perceptions of the purpose of models. Is a model intended to predict future events? Does it make unconditional claims about the consequences of alternative actions? Or, in contrast, is its purpose to aid in making better decisions? Is its purpose to communicate, organize data, or persuade? The way we evaluate models should be consistent with the purpose of the model. We would not demand that a road map display..
topographic features and might not even demand that it be drawn accurately to scale, but the same would not be true of an aircraft navigation map. Likewise, models intended to help decision makers need not pass the tests required of claims to scientific truth.

Our argument starts with the premise that cost-effectiveness models are meant to be aids to decision making. Such models should be used only after careful testing to ensure internal accuracy (internal validity), to ensure that their inputs and outputs are consistent with available data (calibration), and to ensure that their conclusions make sense (face validity). To the extent that different models of the same decision come to different conclusions, modelers should also be expected to explain the sources of the differences (convergent validity). These prerequisites for responsible model use are not controversial. More controversial is to what extent prospective tests of the concordance between model outputs and actual events in the future (predictive validity) are required. The authors argue that tests of predictive validity are valuable but not absolutely essential. The criterion for determining whether, and to what degree, tests of predictive validity are required prior to model use depends on the benefits in terms of improving the model for decision making, and the costs of delaying the flow of information while obtaining the additional data.

The concept of expected value of information (VOI) in decision theory is directly applicable to the assessment of whether more empirical validation is justified. The VOI is the difference between the expected consequences (utility) of a decision guided by a particular piece of information and the expected consequences (utility) of having to make that decision without the information [1]. VOI is also applicable to information that can reduce uncertainty about parameters in a decision, such as the efficacy of a treatment, the duration of a treatment’s effect, or its applicability to different patient groups. Deferring a decision until more information about those parameters is in hand is an option, albeit at a cost of resources, delay, and possibly errors of omission or commission in the interim. Resources may be squandered, or patients harmed, if the more costly or less effective treatment option is used while awaiting definitive evidence. This is the applicable concept for the regulation of models under FDAMA. Would the expected value of more information about the consequences of intervention outweigh the cost of obtaining that information, including the forgone health benefits while awaiting the information? Methods for applying VOI analysis to a clinical decision model have been developed and demonstrated [2].

Whether formally or informally, models are always required to extrapolate evidence through time and space, from a study population to an individual patient, from the end of a clinical trial to the end of a patient’s illness or life, from last year to next year. A policy of banning the use and dissemination of information until perfect evidence is available to support it would paralyze the practice of medicine. There must exist a balance between the costs and consequences of obtaining and waiting for better data and the costs and consequences of permitting a synthesis of the available evidence to influence decisions.

Models are used routinely to guide, or even dictate, public policy decisions in many areas that affect human life and health. Environmental regulation and military planning and strategy are two areas where models have gained stature as policy tools. Beyond these, demographic models are used routinely in the US Bureau of the Census with implications for Social Security benefits, in economic forecasting with implications for macroeconomic policy, in transportation planning with implications for the location and operation of traffic controls and the design of roadways, and in many other areas. As in health care, advocates and critics of models exist in each of these areas. The authors have reviewed the use of models in selected areas of environmental and military decision making with the purpose of drawing lessons for health care from documented experience with models in these fields. We would not claim that just because models are sanctioned, or even required, in other domains of policy that these practices should be applied to health policy. We do argue, however, that there are compelling reasons why these uses of models are appropriate, and that these reasons apply in health policy albeit with important differences.

This paper is organized as follows. Following this introduction, the term model is defined. The authors present evidence that models are ubiquitous, not only in environmental decision making, but in medical decision making as well. In fact, we argue that clinical trials are themselves models, requiring numerous assumptions and extrapolations in order to derive inferences and clinical implications from them. The next section describes the use of models in two areas of environmental policy: registration of new pesticides and regulation of chlorofluorocarbons (CFCs). We also briefly re-
view some uses of models in military decision making. Next, we cite examples of models in health care and public health that have been influential in public policy and in the formulation of practice guidelines. Then we turn to the issue of evaluation of models and propose a taxonomy and conceptual framework for evaluation. Finally, we return to the implications for the implementation of Section 114 of FDAMA and appropriate public policy concerning the use and dissemination of models to guide health-care decisions.

**What Is a Model?**

We all encounter models in our daily lives: road and subway maps are examples. Other models include plastic or electronic architectural representations of cities, the equations of Newtonian and quantum mechanics, or double-helix renditions of the DNA molecule. Two experts on groundwater modeling commented, “... the word *model* has so many meanings and is so overused that it is sometimes difficult to know what one is referring to. ...” [3]. Hereafter, we restrict our attention to mathematical simulation models, as opposed to physical models or other kinds of models. The National Research Council, in its report on the uses of micro-simulation modeling for social policy, offered this definition of a simulation model: “... a replicable, objective sequence of computations used for generating estimates of quantities of concern. ...” [4]. For our purposes, we define a model as an analytic methodology that accounts for events over time and across populations based on data drawn from primary and/or secondary sources.

Box et al. [5] distinguish between empirical and theoretical (or mechanistic) models. An empirical model is used to hypothesize about a situation in which “the mechanism underlying a process is not understood sufficiently well, or is too complicated, to allow an exact model to be postulated from theory.” In a purely empirical model, the data speak for themselves, and there is not a complete, logical story line connecting cause (inputs) and effect (outputs). In contrast, a theoretical model is “based directly on an appreciation of physical or mechanistic theory governing the system.” A single clinical trial can be thought of as an empirical model, whereas a decision-analytic model with empirically estimated parameters, including some estimated from clinical trials, could be considered as a theoretical model. Box et al. cite statistical, data-driven models such as linear regression as examples of empirical models, but they favor theoretical models “whenever a basic understanding of the system is essential to progress.” They contend that a theoretical model (with parameters estimated from data) provides a better basis for extrapolation than does empiricism. The disadvantage is that, although a theoretical model can more closely represent outcomes over a wider span of parameter values than a purely empirical model, extrapolation of either type of model beyond the range of data is never completely safe.

A clinical trial, the epitome of scientific objectivity and direct evidence, is a model. Even if the only item of interest in a study is to learn how an intervention affects a particular study population, modeling is at work: the process of drawing inferences for the closed study population may involve statistical modeling of measurement error or ascertainment bias, curve-fitting, and other manipulations of data. However, clinical trials are not really about the people in the study population. Trials are used, at least implicitly, with the assumption (or model) that the study population is like other people not in the trial, and therefore the conclusions from the trial can be applied or extrapolated to other people. This mapping from a trial population to real-world target populations actually involves several modeled elements, including the following:

- Functional forms of mathematical relationships between exposure and response (e.g., additive, linear, logistic) are specified in order to perform tests of hypotheses and to estimate magnitudes of effects.
- Results from the study population are applied to target populations elsewhere in time and space, often by assuming similarity based on a set of patient characteristics, and often by using statistical models with covariates.
- Surrogate end points in clinical trials are usually used in lieu of truly valued outcomes. Cholesterol-lowering trials that used lipid levels as end points are open to criticism for failing to demonstrate the link to reduced morbidity and mortality. Even trials that use mortality as an end point still require models to assess the consequences in terms of life expectancy gain or quality-adjusted life expectancy. A notable example is the 4S trial of simvastatin to lower cholesterol in persons with prior heart disease; although the trial contained evidence of a reduction in mortality, estimates of the magnitude of the life-expectancy benefit and cost-effectiveness required a model based on external data and judgment [6].
• Observed intervention strategies are interpolated or sometimes extrapolated to optimize dosage or the frequency of intervention in relation to individual patient characteristics. This is especially true in the formulation of screening guidelines, for which direct observations of the range of practiced screening frequencies and modalities are not available within clinical trials. Although the benefits of screening mammmography have been demonstrated experimentally in clinical trials, the incremental benefits of annual mammography have not.

• Economic evaluations performed alongside or in the wake of a clinical trial require modeling, even if the central hypothesis has been experimentally verified. Extrapolation from the trial setting to different localities (or nations) with varying clinical practice patterns, clinical skills, and cost drivers that may or may not coincide with those in the trial protocol, requires the use of models.

In all of these instances, the legitimacy of the assumptions and computations required to translate from empirical evidence to policy is a matter of judgment, and of modeling. Indeed, all inference from data requires modeling, and it is a matter of degree—based on value versus cost of information—that will determine how much data are required to support a model that is intended to inform decisions.

Use of Models in Policy Decision Making

Environmental Models

We have selected two examples of modeling in environmental decision making—pesticide regulation and chlorofluorocarbon regulation—to illustrate the use of models by the US Environmental Protection Agency (EPA) and other regulatory authorities responsible for the control of air, water, solid waste, toxic substances, and nuclear waste pollution. The decision by the EPA to permit a new pesticide on the market and regulate its residues bears some similarities to decisions regarding the approval and marketing of new drugs. In both cases, the policy decision concerns a new product with possible health risks and benefits and possible economic costs and benefits. The consequences of the status quo—not to permit the product to be used or promoted—could be regarded as less risky, or at least better known. However, decisions are often made, based on less-than-ironclad evidence, to permit new pesticides on the market despite uncertainties associated with their use.

Although we focus on the examples of pesticides and CFCs, both of which involve a statutory risk-benefit test by the EPA, models are used throughout the EPA under virtually all of its statutory authorities, even when law does not permit risk-benefit tradeoffs. Under authority of the Clean Air Act, for example, the EPA uses model-based estimates of health risk to support decisions to allow sources of air-pollution to operate [7], despite the possibility—which cannot be ruled out—that the resulting pollution may cause health risks. In such situations, the EPA trusts models that support the abandonment of the purportedly safer status quo.

Pesticides. The EPA uses models in two major aspects of regulation of pesticides: registration and the establishment of tolerances. The details regarding regulation of pesticides may change under the recently enacted Food Quality Protection Act. We describe the process as it has evolved under existing laws, with the expectation that the role of models will not materially change under the new law.

For every intended use of a pesticide (i.e., the application of a pesticide on a particular crop), the pesticide manufacturer must register with the EPA. The EPA does not require proof of safety of the pesticide prior to registration. Rather, the EPA is authorized to use models to assess human health risks and arrive at a judgment as to whether the benefits of use outweigh these risks. Data to support these models may consist of up to 70 kinds of laboratory and field tests, at a cost of up to $10 million over a period of 6 to 9 years, for each intended use. If pesticides are used on food crops, then the EPA must also establish residue tolerance for each particular use. Tolerances are legal limits on the amount of pesticide residues that may be present on crops sold in commerce. They are proposed by the manufacturer and are required to be high enough to allow for effective application without being exceeded, but low enough not to pose unreasonable dietary risks.

For the purposes of both registration and tolerance determination, models are used by the EPA in three ways: to estimate quantities of dietary exposure to pesticide residues; to assess health risks in relation to dose (using mostly animal models); and to estimate concentrations of residues that will end up in groundwater, rivers, and streams and the consequent implications in terms of human exposure and risks. This information is combined into a calculation of the estimated excess lifetime risk of cancer in the affected population. As per EPA regulatory policy, a tolerance is ap-
proved if the modeled estimate of risk is less than one in 1 million, a risk-benefit analysis is performed if the estimate is between one in 1 million and one in 10,000, and the tolerance is rejected if the estimate exceeds one in 10,000.

Exposure models used by the EPA must estimate both the quantities of residue that will remain on food when it is consumed (after processing and washing) and the quantities of treated foods that are consumed. The EPA uses a computer-based simulation model that requires data about the types and amounts of food that people consume and on the pesticide residue levels [8]. These input data are combined to calculate exposure for specified subpopulations defined by age, gender, ethnicity, geography, and season. EPA calculates dietary exposure to pesticide residues in food by using the concept of a Theoretical Maximum Residue Contribution (TMRC), which reflects an upper bound (or worst-case) scenario for exposure. In calculating the TMRC, EPA estimates average food consumption based on the National Food Consumption Survey and assumes that residues are equal to the tolerance level and that 100% of the crop in question is treated with the pesticide. The model estimates the amounts of raw ingredients used in prepared foods using standard recipes [9]. Use of the TMRC for exposure assessment is likely to overstate risk as a result of washing, processing, and incomplete application of the pesticide. If a preliminary analysis based on TMRC shows that the resulting exposure estimates are high enough to be of concern, the model is rerun with more realistic assumptions called Anticipated Residue Concentrations (ARC), which reflects an upper bound (or worst-case) scenario for exposure. In some cases, the EPA might require the registrant to collect data on residues in food ready for consumption (i.e., market-basket studies).

Some critics complain that conservative estimates of risk may be interpreted as actual risks [10], and others worry that the special diets of infants and children may not be reflected adequately by the aggregate models. Despite the important limitations of the EPA exposure model and its required data inputs, these limitations do not preclude its use in the determination of regulatory policy.

The second step in risk analysis for pesticides concerns the relation between exposure and human risk. Human risk assessment models used by the EPA accept data from human epidemiology, animal experiments, in vitro studies, and theoretical considerations based on molecular structure.

Although epidemiology has played an important role in human health-risk assessment [11], it has clear limitations when it comes to toxicity assessment of environmental contaminants such as pesticides [12]. Epidemiological studies have poor sensitivity because they are unable to detect low risks. They also have poor specificity, because of their inability to distinguish and eliminate other confounding causes of disease. Problems of confounding, population heterogeneity, and long lag times between exposure and effect all conspire to limit the usefulness of direct human evidence. For this reason, the EPA does not consider the absence of epidemiological data sufficient grounds for denying the registration of a pesticide (though limitations of epidemiological data do not preclude their use).

Instead of human data, EPA models rely mostly on animal experiments to estimate dose-response relationships. One of the strengths of this type of data is that they are randomized and therefore offer some protection against confounding. Limitations are that models are required to extrapolate from very high doses to those that would be expected to reflect exposure in a real-life situation such as at the dinner table, and to extrapolate from rodents to humans [13]. The EPA has developed a series of risk assessment guidelines for modeling carcinogenic, reproductive, and developmental risks. Empirical justification for these assumptions is limited, but they are used because they are regarded as conservative and because it would be prohibitively costly, if not impossible, to verify them.

Estimation of the risk associated with the fate and transport of pesticide residues is problematic. Although the EPA does sometimes require field-scale experimental trials, it does so infrequently because of the high cost, uniqueness, and lack of generalizability of each experimental setting, and the variability of weather conditions. Instead, retrospective data collected from monitoring stations on streams and rivers and in groundwater are used in computer simulation models to estimate future ambient concentrations in water [14,15].

Thus, the EPA uses a stepwise approach to requiring data to support its decisions. More data are required if and only if the model indicates that the decision would be sensitive to the findings. This is none other than the value-of-information principle, which is equally applicable in the context of FDA policy regarding pharmacoeconomic models.

**Chlorofluorocarbons.** CFCs are very stable chemicals that were developed in the 1930s and widely used as refrigerants, as blowing agents in manufac-
turing insulating and packaging foam, and as solvents for dry cleaning, microelectronics, and other uses. CFCs first attracted concern when it was reported that they could reduce the concentration of stratospheric ozone, which is the primary agent that screens out excess ultraviolet-B (UV-B) radiation before it reaches the Earth’s surface [16]. Increased human exposure to UV-B radiation promotes skin cancer and cataracts and has adverse effects on various crops and wildlife.

Manufacture and use of CFCs were ultimately banned on a global scale under the terms of the 1987 Montreal Protocol and subsequent amendments. International agreements and domestic regulations have been crafted that rely almost entirely on theoretical predictions derived from mathematical simulation models. This reliance on models was widely accepted, at least in part because it was understood that CFCs would persist in the atmosphere for at least a century. If one waited for definitive empirical evidence of ozone depletion, let alone evidence of increased incidence of skin cancer and cataracts, before reducing atmospheric emissions, it might require a century or more to reverse any harmful effects. These decisions were made in a context in which the economic costs and health risks of substitute products were acknowledged.

In the United States, the first regulatory action was the 1979 ban on the use of CFCs as propellants in aerosol dispensers for hair spray, deodorant, and other personal-care products. These uses accounted for about half the CFC consumption in the United States at the time and were regarded as relatively low-value uses. As research on the effects of CFCs continued, model-based projections of the resulting magnitude of ozone depletion oscillated, although continuing to support the basic hypothesis. An early model projected that if CFC emissions continued at 1974 rates, then stratospheric ozone would stabilize at 7% to 13% below the baseline level [16]. A series of four National Academy of Science reports, which examined the current scientific theory, laboratory experiments, and simulation-model results, reported model-based estimates of ozone depletion ranging from 2% to 16% [17–20]. The World Meteorological Organization 1985 summary report showed that uncertainties in simulation-model coefficients resulted in a range of equilibrium depletion from less than zero to more than 20%, with even wider ranges associated with higher, more realistic emission rates [21].

With this history of varying model projections as virtually the sole quantitative basis for decision making, the Vienna Convention was signed in 1985. The Convention provided an international framework for cooperative study and possible restrictions on CFC use. International limits on CFC consumption were specified by the Montreal Protocol and signed by most of the large, industrial nations in September 1987.

The first direct observational evidence of stratospheric ozone depletion became available between the signings of the Vienna Convention and the Montreal Protocol. In May 1985, the British Antarctic Survey reported that it had observed temporary but substantial depletion of ozone above its Antarctic research station each spring beginning in 1981 [22]. There was, however, no specific link to CFCs, and the possibility that CFCs could induce such a localized effect had not been previously recognized. Indeed, for the next several years scientists worked to determine whether the hole was caused by CFCs or by dynamic mechanisms involving sinking of low-ozone air from higher elevations. The actual cause, suggested in 1986, was attributed to surface reactions on polar stratospheric clouds which inactivate chemicals that would otherwise remove chlorine and thus prevent ozone depletion [23], and preliminary measurements supporting this hypothesis became available at the time of signature of the Montreal Protocol. Definitive evidence supporting the role of CFCs in creating the hole did not become available until 1988, after the Protocol was signed. Statistical analysis of satellite measurements provided empirical evidence of a downward trend in global ozone concentrations beginning in March 1988 [24]. Despite the health risks and economic costs of the alternatives to CFCs, the critical decision to limit their use was made on the strength of mathematical models rather than direct evidence.

**National Defense**

Modeling to guide military decisions is so pervasive and established that some of the most thoughtful, scholarly writing on the role of modeling in policy making has come from this field [25]. Models are routinely used for logistical planning, but more germane to our topic of life-and-death decisions, simulation models are used to guide decisions about the choice of weapon systems and even to guide combat decisions. Despite the acknowledged limitations of these models, they are used to guide the planning, deployment, and use of the second largest commitment of human resources in the federal budget after health care.


Health Policy
It is useful to examine how models have been used to influence resource allocation decisions in health care. Our purpose is only to show that public-sector and private-sector health-care decision makers do pay attention to models. The implication is not necessarily that the models lead to better decisions—a question that we will return to later—but that they are perceived as valuable by organizations entrusted with our health-care dollars. We give examples of models that have been influential in public health policy, coverage decisions by public and private insurers, and the formulation of practice guidelines.

The US government has a long history of using and developing models to guide public health policy. Models, rather than direct evidence of cost-effectiveness, have supported vaccine recommendations by the Centers for Disease Control and Prevention (CDC) at least since the late 1960s. The CDC’s report “An Ounce of Prevention” [26] reviewed and endorsed a variety of model-based estimates of gains in quality-adjusted life expectancy and cost-effectiveness ratios. The CDC also used a model to guide its decision concerning screening for thyroid disease in persons exposed to radioactive iodine (I-131) near the Hanford (WA) nuclear weapons facility. Its recommendations that pregnant women increase their consumption of folate-rich foods to prevent neural tube defects were based on a model that synthesized evidence from a variety of sources, rather than on direct evidence of benefit. As another example of government reliance on models, the National Institute of Allergic and Infectious Diseases commissioned the Institute of Medicine twice—with reports issued in 1985 and again in 1999—to use models to recommend priorities for the development of new vaccines [27,28].

The US Health Care Financing Administration based its decisions to cover immunization against pneumococcal pneumonia (1981) and influenza (1993) under Medicare on influential models [29,30]. Subsequently, Medicare has appealed to strong but indirect evidence of cost-effectiveness from models to cover magnetic resonance angiography in lieu of carotid angiography prior to carotid atherectomy, and to cover erythropoietin as adjuvant therapy under the end-stage renal disease program (personal communication: Sheingold S, US Health Care Financing Administration). Evidence supporting the cost-effectiveness of pharmaceuticals is now required in support of coverage decisions in Europe, Canada, and Australia, and guidelines for the use of models have been promulgated. The Academy of Managed Care Plans and several individual managed care companies in the United States have also issued guidelines for models. Of 70 cost-effectiveness evaluations submitted to the Australian Pharmaceutical Benefits Advisory Committee between December 1995 and June 1997, only 17 were based exclusively on direct evidence from randomized clinical trials, 24 were based only on models, and 29 used a combination of direct experimental evidence and modeling (personal communication: Mitchell A, Australian Pharmaceutical Benefits Advisory Committee). Likewise, several model-based studies submitted to the Canadian Coordinating Council of Health Technology Assessment (CCHOHTA) led to coverage recommendations, including a decision to cover the drug omeprazole to treat gastro-esophageal reflux disease (personal communication: Otten N, formerly with CCHOHTA). In managed care, the role of models—and of economic evaluations in general—remains uncertain, but decisions to cover or not to cover such technologies as enhanced cervical cytology screening and BRCA gene testing in various health-maintenance organizations appear to have been guided in part by models (personal communication: Eddy D, MD, PhD, consultant).

The role of models in influencing the formulation of influential clinical practice guidelines is clear. The cervical cancer-screening model developed by Dr. David Eddy—in the absence of direct experimental evidence of benefit—was cited in the American Cancer Society’s revision of its cervical cancer-screening guidelines. That model suggested that the incremental value of annual screening after consecutive negative screens was negligible, and that less frequent screening, such as every 3 years, was cost-effective [31]. Likewise, an analysis based on the Coronary Heart Disease (CHD) Policy model [32] was cited by the Second Adult Treatment Panel of the National Cholesterol Education Program to support its recommendation strongly endorsing treatment of mildly elevated low-density lipoprotein (LDL) levels in persons with a history of heart disease [33]. This recommendation was made even before direct evidence from clinical trials confirmed that drug-mediated lipid lowering reduced the risk of CHD mortality and nonfatal events by approximately the degree predicted by the model.

Model Validation and Evaluation
Many diverse notions of validation or validity have been applied to policy models. Users and
critics of policy models believe that some sort of validation is necessary to protect against incorrect conclusions and to protect users from being misled by models. Terms and concepts related to validation, albeit with somewhat different meanings, include verification, face validity, and corroboration. Here, we will use the term validity to refer to the more specific concept of predictive validity, i.e., whether a model produces outputs that are consistently and reliably borne out by actual events. We use the term verification to denote a weaker condition, i.e., that the model’s inputs and outputs are consistent with actual events, laboratory findings, or generally accepted theories. Included within the process of verification is what is often called calibration, i.e., ensuring that the model’s outputs are consistent with known data at the aggregate level. Also included within the process of verification is the process known to modelers as debugging, or ensuring that the inner workings of the model are behaving as intended and are free of coding errors and logical inconsistencies. Face validity concerns whether the outputs of a model make sense and can be explained intuitively. We use the term corroboration synonymously with convergent validity, i.e., whether two or more independent models or studies lead to the same or similar conclusions, or, if they do not, whether the differences can be explained. Finally, we use the term evaluation to encompass all of the above notions of validity. Evaluation is our preferred term, because it is the most general and does not impose any particular viewpoint regarding expectations for policy models.

Based on a review of the literature on policy models and on careful consideration of their purposes, we conclude that establishment of predictive validity is often valuable, but not a prerequisite for model use. Some models may be impossible to validate in this stringent sense. Verification and calibration are intended to demonstrate that a model can mimic past behavior, whereas predictive validation refers to assessing whether a model can predict the future [34]. The EPA Task Force on Environmental Modeling supported the more limited criterion for evaluation, saying that “Models are never literally validated, instead they are invalidated” [35]. In statistical inference, one never proves the null hypothesis; one can only disprove it with reasonable certainty. The same is true of models. If the predictions of models are consistently contradicted by facts, then the model can reasonably be rejected—and possibly revised for future use. Otherwise, use of the model can be said to be consistent with, or verified by, reality.

**Model Verification (Calibration)**

Verification of a model consists of demonstrating that its inputs and outputs are consistent with known facts, and that it is functioning properly in a technical sense. Verification includes tests of both the internal and external consistency of the model with known facts.

The process of testing for internal consistency is called debugging by experienced modelers. In debugging, a model is often subjected to various extreme input conditions to test whether it gives the expected outputs. For example, if a disease-specific mortality rate is set to zero, then the number of deaths caused by the associated disease should be zero. As another example, if the efficacy of a treatment is zero, then health outcomes should be identical with or without that treatment. Savvy modelers know how to debug thoroughly and convincingly, and this type of verification of extreme cases is much more effective and efficient than staring at computer code looking for typographical or logical errors.

Calibration of a model against real data is the next step in verification, and entails testing for consistency with observed reality. For example, a model of a disease that incorporates estimates of incidence, progression, and survival rates based on clinical data must then be tested against independent estimates of aggregate numbers of cases and deaths from the disease, such as might be obtained from national vital statistics or population-based health services utilization data. An iterative process ensues, in which different interpretations of the input and output data are tested in the model until consistency is achieved with both. A description of the calibration process should be included in any report about a model conveyed to decision makers.

**Convergent Validity**

Convergent validity of models concerns the ability of independently developed models to give similar results. Convergent validity tests for health-care models were recommended in the report of the Panel on Cost-Effectiveness in Health and Medicine as a response to the problem that the Panel called model process uncertainty [36]. Of course, convergent validity tests may not be helpful if models are not truly independent, i.e., if they are all built on the same flawed assumptions. Thus, convergent validity tests can be regarded as a source of increased confidence in models, although they cannot overcome deficiencies in inputs or logical assumptions.
Face Validity

Although it is inappropriate to claim that face validity is sufficient for model use, its absence should cause decision makers to be skeptical. Relying exclusively on face validity can be hazardous if measurements have not or cannot be made. For example, until this century, Newtonian mechanics had face validity because their predictions were consistent with available measurement techniques. Special relativity could be shown empirically to be a superior model only when technology improved to the extent that velocities close to the speed of light could be measured. An example in health care relates to models of the safety of the blood supply, which once held that the risk of transmission of the human immunodeficiency virus (HIV) increases in proportion to the prevalence of infection in the donor population. More sophisticated models that incorporated seroconversion claimed that the risk of transmission could actually be higher in lower-prevalence populations in which the incidence of infection was growing rapidly [37]. Until data became available to support the newer model, their face validity was widely questioned, and blood-screening policies continued to be based on the incorrect model.

Predictive Validity

The prevailing view of model validity in public policy is summarized in an essay by two experienced modelers in the defense arena: “Only models intended to predict need to be validated. [. . .] Some models can be validated and used to predict, and others cannot be validated and may be put only to nonpredictive uses.” [25] If a model is intended to help decision makers, and not necessarily to predict the future, then it is inappropriate to demand predictive validity. This is not to say that prospective validation of models is not valuable, but the value of the information obtained must be weighed against the costs of obtaining it.

When can models be truly validated in the predictive sense? Hodges and Dewar [25] list four necessary conditions that a situation must satisfy for predictive validation to be a reasonable goal. First, the situation must be observable and measurable. Second, the situation must exhibit constancy of structure over time. This condition would fail to hold, for example, in a premature or static attempt to validate a model of the clinical performance of a diagnostic technology whose users gain skill over time through experience. Third, the situation must exhibit constancy across variations of conditions not specified in the model; this is necessary in order to be able to validate a model under a wider range of conditions than those for which data can be obtained. Fourth, the situation must permit the collection of ample data with which to make predictive tests of the model. As Hodges and Dewar observe, “Many models [. . .] cannot be validated, so it is pointless to try.” [25]

The situations for health-care models would seem typically to defy these criteria for predictive validation. Circumstances do change over time, so that the conditions that held when the model was designed may not apply in the future. Models developed for one setting cannot be validated predictively in other settings if the variables that differentiate the settings are not explicit within the models. Moreover, models designed to predict the incidence and consequences of unique events, such as epidemics, cannot be validated predictively because of insufficient data points. Does the inability to validate a model predictively render the model useless? Hodges and Dewar are emphatic in their response to this question: “Few military models or models of human decision making can be validated, and it is counterproductive to demand as a matter of policy that users and institutional parents attempt to validate them.” They go on to state: “We reiterate that a model that cannot be validated in [the predictive] sense is not necessarily useless; it simply may not be used to make sentences like ‘the model says X.’”[25]

The implications of this view of validation for the regulation of pharmacoeconomic claims by the FDA are clear: As long as unvalidated predictive statements of the type “Drug A will lead to outcome X” are avoided, models can be useful aids to decision making if their claims are stated as contingent upon their assumptions and data inputs. Furthermore, models intended for nonpredictive uses, including uses as decision aids, can be evaluated successfully using other criteria such as verification, corroboration, and face validity.

Models whose outputs are framed as point estimates cannot usually be validated in the predictive sense because they provide no metric for determining whether a prediction is close. It may be possible to establish external standards to determine whether a prediction is close enough for the decision-making purpose intended, but these may be viewed with suspicion unless the standards are set in advance of the validation exercise. Models whose outputs are framed probabilistically offer greater opportunity for probabilistic validation. For example, weather forecasting models can be, and are, validated when their predictions are framed in proba-
bilistic terms (e.g., the probability of rain is 40%). By studying the frequency of rain on days for which the model claims a 40% probability, the possibility exists to compare the actual and predicted frequencies of rain. By analogy, health-care models that produce estimates of uncertainty around model predictions—as from Monte Carlo simulations and stochastic sensitivity analyses—can be validated predictively provided that the outputs of the model are measurable. Even if the outputs of a health-care model are not measurable, inputs may be subject to validation if the modelers provide uncertainty bands around them. For example, a model of a cholesterol-lowering intervention in the prevention of heart disease could be said to be validated if empirical evidence of the quantitative association between cholesterol changes and heart disease events fall within the range of parameter estimates provided with the model. The National Research Council’s report on microsimulation modeling espoused this view of validation as a process for measuring the uncertainty or variability in a model’s estimates and identifying the sources of that uncertainty [4].

Tests for predictive validity effectively expand upon tests for verification by expanding the set of observations against which model results can be verified. The decision to embark on, or require in a regulatory sense, additional data against which to validate a model ought to be guided by the value-of-information principle; the value of reducing uncertainty in the model in terms of the expected outcomes of improved decisions must outweigh the cost of obtaining, and waiting for, the information.

**Approaches to Model Acceptability and Accreditation**

Recognizing that true validation of models is generally impossible and, as some authors have noted, useless to attempt, alternative criteria based on whether the model is acceptable to its users, or accredited by independent adjudicators, may be invoked. Such criteria for model evaluation have been espoused in the environmental and defense areas, and they have direct applicability to health-care models. The Pharmaceutical Benefits Advisory Committee in Australia requires a stringent accreditation process for all models submitted as evidence of cost-effectiveness, including a peer review that often catches technical errors [38].

Because it relies on models to support regulatory decision making, the EPA has devoted considerable resources toward developing criteria for the acceptability of models. In response to the growing need to establish general criteria for model evaluation, an Agency Task Force on Environmental Regulatory Modeling (ATFERM) was established in 1992 and reported its findings in 1994 [35].

The ATFERM identified four criteria for evaluation of models: appropriateness, accessibility, usability, and reliability. Appropriateness refers to whether the model is appropriate for the physical system being described. This criterion asks whether the logical assumptions and relationships among variables are consistent with theoretical and empirical evidence about the physical system, be it transport of toxins through air or water, or human biologic response. Accessibility examines whether the model is available for public review. This criterion was echoed by the Panel on Cost-Effectiveness in Health and Medicine in its recommendations regarding reporting of cost-effectiveness analyses [36]. Models should be open to peer review, possibly under controlled and confidential conditions that protect proprietary interests in the code. Usability is concerned with functional and operational aspects such as the cost and feasibility of running the model in a timely fashion.

It is apparent that actual application of these criteria requires a certain amount of subjectivity and interpretation, but the EPA has determined that it is within its purview to make those judgments and thereby permit the use of models that it judges acceptable. Moreover, the EPA’s judgment on acceptability is based in large part on the model’s acceptability to its peer reviewers and users.

The defense modelers, Hodges and Dewar, propose criteria for evaluating models used for marketing defense-related products, and these criteria have clear analogies to the regulation of models designed to sell pharmaceutical products. They caution that models, especially the simple models often used for marketing purposes, could be inappropriately used to make predictions and that the modeler is responsible for disclaiming any intention to do so. The forecasts should be framed as contingent upon assumptions and not as unconditional predictions. They conclude that “the analystscum-salesman is not off the hook: he must have a good idea about how to produce the benefits and must accompany presentations with appropriate caveats.” [25]

Hodges and Dewar also comment on what they call a *forfìtori* arguments, or what is often referred to in health-care modeling as worst-case assumptions, or bending over backwards. In other words, the modeler makes extreme assumptions unfavorable to the conclusion and is still able to reach the..
results are stated as conditional upon input assumptions, tests of verification, such as debugging and calibration, could be regarded as sufficient. Transparency is essential in describing the input values, the logical assumptions, and the process of verification. In keeping this perspective, the input data themselves need not be validated so long as they are not stated as claims and so long as the results of the model are stated conditionally upon the inputs.

Realistically, this standard of stating claims as conditional upon assumptions may be difficult to enforce. Therefore, a policy that recognizes the value of, and need for, prospective data collection under certain circumstances is needed in order to guard against relying on unsubstantiated assumptions. Section 114 of FDAMA defines a criterion of competent and reliable scientific information as a basis for economic claims. According to the US Senate and House Reports on FDAMA, the intent of Congress was to apply the definition of competent and reliable scientific information that had been previously articulated and implemented by the Federal Trade Commission (FTC) in regulating advertising claims [43]. The key to the FTC standard, in turn, rests on its definition of prior substantiation, namely, the degree of evidence in support of the claim. The FTC does not require absolute proof in support of a claim. Rather, it applies a cost-benefit test to the process of collecting information in support of the claim. According to the FTC regulation, what defines reasonable substantiation depends “on a number of factors relevant to the benefits and costs of substantiating a particular claim.” [43] The regulation lists among these factors: “the consequences of a false claim, the benefits of a truthful claim, and the cost of developing substantiation for the claim.” [43]

As this article is being written, there remains some dispute as to whether the competent and reliable standard in Section 114 of FDAMA applies both to the effectiveness and to the cost estimates in a cost-effectiveness model. Logically, the same criteria for weighing the value and cost of additional information to support decisions affecting the cost-effective use of drugs ought to be applied to all information that is relevant to the decision maker. However, the FDA currently interprets the statute to restrict the dissemination of models that incorporate estimates of clinical effects outside the approved indications. This interpretation could have the effect of impeding the flow of information to decision makers, even if the conclusions from the models are clearly stated as contingent upon the assumed extrapolations beyond proxi-

Validation of Health Care Decision Models

If one accepts the conditions stated by Hodges and Dewar, then many health-care models would not be susceptible to true predictive validation. Generally, situations being modeled are changing in ways not explicitly included in the model. Technologies being modeled, and those to which they are being compared, may change between the time the model is developed and the time data become available. Population characteristics may change as a result of environmental changes or changes in a disease entity. A striking example of the latter was the onset of resistant strains of HIV after initiation of use of antiretroviral therapies such as zidovudine [41], thereby invalidating the previous model [42].

Although it should not be a prerequisite for model use, predictive validation can be valuable. It can help modelers revise models for future use, in effect extending the range of data against which to calibrate them. It can also help users of models guard against basing future decisions on models that have been found to be inconsistent with known data, although it need not deter them from using revised versions of such models that have been verified.

Implications for Implementation of FDAMA

The fundamental policy question facing the FDA under FDAMA is under what conditions to permit drug manufacturers to use models as the basis for claims of cost-effectiveness. As long as modeling results are stated as conditional upon input assumptions, tests of verification, such as debugging and calibration, could be regarded as sufficient. Transparency is essential in describing the input values, the logical assumptions, and the process of verification. In keeping this perspective, the input data themselves need not be validated so long as they are not stated as claims and so long as the results of the model are stated conditionally upon the inputs.

Realistically, this standard of stating claims as conditional upon assumptions may be difficult to enforce. Therefore, a policy that recognizes the value of, and need for, prospective data collection under certain circumstances is needed in order to guard against relying on unsubstantiated assumptions. Section 114 of FDAMA defines a criterion of competent and reliable scientific information as a basis for economic claims. According to the US Senate and House Reports on FDAMA, the intent of Congress was to apply the definition of competent and reliable scientific information that had been previously articulated and implemented by the Federal Trade Commission (FTC) in regulating advertising claims [43]. The key to the FTC standard, in turn, rests on its definition of prior substantiation, namely, the degree of evidence in support of the claim. The FTC does not require absolute proof in support of a claim. Rather, it applies a cost-benefit test to the process of collecting information in support of the claim. According to the FTC regulation, what defines reasonable substantiation depends “on a number of factors relevant to the benefits and costs of substantiating a particular claim.” [43] The regulation lists among these factors: “the consequences of a false claim, the benefits of a truthful claim, and the cost of developing substantiation for the claim.” [43]

As this article is being written, there remains some dispute as to whether the competent and reliable standard in Section 114 of FDAMA applies both to the effectiveness and to the cost estimates in a cost-effectiveness model. Logically, the same criteria for weighing the value and cost of additional information to support decisions affecting the cost-effective use of drugs ought to be applied to all information that is relevant to the decision maker. However, the FDA currently interprets the statute to restrict the dissemination of models that incorporate estimates of clinical effects outside the approved indications. This interpretation could have the effect of impeding the flow of information to decision makers, even if the conclusions from the models are clearly stated as contingent upon the assumed extrapolations beyond proxi-
mal end points in clinical trials or to other target populations.

The decision-theoretic concept of VOI can be used to address the cost-benefit test implied by the “competent and reliable” standard [2]. VOI measures the difference between the expected consequences of making a decision with additional information and the expected consequences of having to make that decision without the information. Embedded within VOI are the relative consequences of errors of omission and commission.

To enable an assessment of the value of information that might solidify the assumptions and input estimates in models, probabilistic representation of model inputs and outputs is required. This is a deficiency of most current models, understandably so because the computational challenges of fully probabilistic modeling are only beginning to be met by technology. But as the new century begins, prospects for practical approaches to probabilistic modeling are emerging that promise to make VOI analysis a practical tool for regulating the flow of information. The implications extend beyond the regulatory functions of the FDA and into the resource allocation decisions of the National Institutes of Health and of private and corporate investors in health research and development.

Conclusions

Mathematical models are used routinely and with formal governmental authorization in environmental and defense-related decision making. Although empirical validation of the predictions of these models is often prohibitively expensive or even impossible, their conclusions are disseminated widely and form the basis for decisions that involve substantial resource commitments and health consequences. For the most part, environmental and military models do not claim predictive validity; rather, they are intended to synthesize current knowledge in a form that can guide decisions. Decision makers in these domains acknowledge the limitations of models but rely on them nevertheless.

It is unclear why models are more widely accepted in other domains of public policy than in clinical medicine. The life-and-death implications of an ill-advised decision to register a pesticide can be at least as dire as the consequences of substituting one safe and effective drug for another. An obvious hypothesis is that modeling is necessary in environmental health because human experiments are physically and ethically impossible, and hence it is accepted as a necessary evil. However, that does not explain why EPA rules about substantiating models with data on exposures or toxicities for new pesticides are more relaxed than FDA rules on pharmacoeconomic claims for new drugs. Perhaps the reason is that models are relatively new in medicine and more established in environmental science. In any case, we believe there are lessons to be learned from all of these areas of model-guided decision making.

Many, if not most, pharmacoeconomic models are also intended to aid decision making rather than predict the consequences of interventions. Models that do not claim predictive validity should not be subjected to the same tests as models that make predictive claims. Modeling results that are stated as conditional upon input assumptions could be considered admissible so long as the logic, code, and integrity of the description of the assumptions and data of the model itself have been verified. Following this perspective, the inputs that enter the model, and the outputs that it produces, need not be validated prospectively so long as they are not stated as claims, and so long as the results of the model are stated conditionally upon the inputs. However, in many circumstances validation of model predictions against prospective data may be valuable in revising the model for future use. The value of such validation must be weighed against the cost.

Decision models that are used for communication with decision makers must be evaluated as to whether the contingent claims they make are correct. Evaluation criteria should therefore address the following:

- Transparency: Are the assumptions and input parameters, and the logic connecting them to outputs, stated with complete clarity and are they open to peer review?
- Verification: Are the outputs of the model consistent with observed data? Has the model been debugged and tested for internal consistency?
- Corroboration: Have other models of the same problem produced similar results contingent upon similar assumptions and input parameters?
- Face validity: Do the results of the model make sense in relation to theoretical considerations, and can they be explained in intuitive terms?
- Accreditation: Has the model been subjected to peer review by a dispassionate reviewer and found to be what it claims to be?

Tests of corroboration and face validity may not always be possible or even desirable, espe-
cially for models whose results contradict the conventional wisdom. These might be regarded as generally desirable, but not necessary, aspects of model evaluation.

Tests of predictive validity of models may or may not be worth the cost, depending on the circumstances. This statement applies to the question of whether direct evidence from clinical trials is required to validate the assumptions or results from models. Decisions and regulations about what empirical and experimental evidence to collect should be guided by the application of VOI analysis to models. It would be preferable to apply the concept of VOI, even if a formal analysis is impractical, than to apply a rigid criterion that prospective data are or are not always required to validate models predictively. In the context of economic models for pharmaceuticals, the test adopted in Section 114 of FDAMA requires that the value of additional data outweigh the costs and consequences of obtaining them. This implies that models for which the cost of obtaining additional information, including forgone benefits from withholding potentially beneficial or cost-effective interventions, is not outweighed by the value of the information, should be considered appropriate for decision making in their present form. When the cost of obtaining the information is justified by the value of information, the model can still be useful for interim decision making, but unconditional claims based on the model should be avoided.

Models are useful guides to decision making, even though they can always be improved with additional time and resources. Whether the time and resources to obtain further validation are justified depends on the value and cost of obtaining the data. It is inappropriate to expect—or demand—that models predict the future accurately, because they can only incorporate what is known at the time the model is brought to bear on a decision. We conclude with a summary comment by the eminent statistical scientist, George Box [5]: “All models are wrong, but some are useful.”

This project was supported by a research grant from the International Federation of Pharmaceutical Manufacturers’ Associations. An earlier version of this paper was presented by Peter J. Neumann, ScD, at the Fourth Annual Meeting of the International Society for Pharmacoeconomics and Outcomes Research, Crystal City, VA, May 26, 1999.

References


