A New Statistical Method to Determine the Degree of Validity of Health Economic Model Outcomes against Empirical Data

Corro Ramos, Isaac; van Voorn, George A K; Vemer, Pepijn; Feenstra, Talitha L; Al, Maiwenn J

Published in:
Value in Health

DOI:
10.1016/j.jval.2017.04.016

VALUE IN HEALTH 20 (2017) 1041–1047

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

A New Statistical Method to Determine the Degree of Validity of Health Economic Model Outcomes against Empirical Data

Isaac Corro Ramos, PhD[1,*], *George A.K. van Voorn, PhD*[2], *Pepijn Vemer, PhD*[3,4], *Talitha L. Feenstra, PhD*[3,5], *Maiwenn J. Al, PhD*[6]

[1]*Institute for Medical Technology Assessment, Erasmus University Rotterdam, Rotterdam, The Netherlands;* [2]*Biometris, Wageningen University and Research, Wageningen, The Netherlands;* [3]*Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands;* [4]*Pharmacoepidemiology and Pharmacoeconomics (PE2), Groningen University, Groningen, The Netherlands;* [5]*National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands;* [6]*Institute of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, The Netherlands*

ABSTRACT

Background: The validation of health economic (HE) model outcomes against empirical data is of key importance. Although statistical testing seems applicable, guidelines for the validation of HE models lack guidance on statistical validation, and actual validation efforts often present subjective judgment of graphs and point estimates. **Objectives:** To discuss the applicability of existing validation techniques and to present a new method for quantifying the degrees of validity statistically, which is useful for decision makers. **Methods:** A new Bayesian method is proposed to determine how well HE model outcomes compare with empirical data. Validity is based on a preestablished accuracy interval in which the model outcomes should fall. The method uses the outcomes of a probabilistic sensitivity analysis and results in a posterior distribution around the probability that HE model outcomes can be regarded as valid. **Results:** We use a published diabetes model (Modelling Integrated Care for Diabetes based on Observational data) to validate the outcome "number of patients who are on dialysis or with end-stage renal disease." Results indicate that a high probability of a valid outcome is associated with relatively wide accuracy intervals. In particular, 25% deviation from the observed outcome implied approximately 60% expected validity. **Conclusions:** Current practice in HE model validation can be improved by using an alternative method based on assessing whether the model outcomes fit to empirical data at a predefined level of accuracy. This method has the advantage of assessing both model bias and parameter uncertainty and resulting in a quantitative measure of the degree of validity that penalizes models predicting the mean of an outcome correctly but with overly wide credible intervals.

Keywords: decision making, health economics methods, statistics, validation.

Introduction

The decision making around the reimbursement of newly developed drugs often involves cost-effectiveness analyses underpinned by health economic (HE) decision models [1,2]. HE decision models, like all simulation models, require validation to ensure the credibility of their outcomes [3]. Validation may be described as "the act of evaluating whether a model is a proper and sufficient representation of the system it is intended to represent, in view of a specific application" [4]. A model that is in accordance with what is known about the system is said to be "proper," and a model whose results can serve as a solid basis for decision making is said to be "sufficient." Models that have not been properly validated could deliver invalid results, and hence lead to biased decisions in drug reimbursement or other areas of health policy applying the results of HE decision models.

Different guidelines for the validation of HE models can be found in the literature, but these are not very specific about the operationalization of validation efforts [3,5]. The validation assessment tool AdViSHE adds to these guidelines by being a tool for structured reporting on all relevant aspects of validation (conceptual model, input data, implemented software program, and model outcomes) but does not indicate any particular methodology [4]. In this article, we provide further details on one of these aspects: the validation of HE model outcomes against empirical data. When the empirical data are not used to estimate the input parameters of the model, this is often called independent or external validation. Otherwise, the validation is called dependent or internal [5]. Although statistical testing seems applicable to assess the validity of HE model outcomes in a setting of uncertain observations of possibly variable outcomes, actual applications often present comparisons in an informal way involving subjective judgment of graphs and point estimates [6–9].

In statistics, accuracy is defined as the combination of (lack of) bias and variance [10]. Bias is the difference between the expected value of the outcome predicted by a model and its actual empirical value. Variance is any measure of variability of a

* *Address correspondence to:* Isaac Corro Ramos, Institute for Medical Technology Assessment, Erasmus University Rotterdam, Room J8-27, P.O. Box 1738, Rotterdam 3000 DR, The Netherlands.

E-mail: corroramos@imta.eur.nl.

1098-3015$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2017.04.016

model outcome. This is a broader concept and does not necessarily refer to the statistical variance of a set of data or a random variable. In prediction modeling, it is common to talk about bias-variance trade-off when discussing prediction errors [10]. In general, bias is reduced and variance is increased as more parameters are added to a model. Large variance means that relatively large differences are observed in the model outcomes with little additional input data. The "classical" trade-off problem consists of exploring different levels of model complexity and choosing the one minimizing the overall prediction error [10]. Nevertheless, in the setting of HE model validation and in this article, the assumption is made that the simulation model is a given and it is investigated whether the current levels of bias and variance are acceptable for decision-making purposes.

Furthermore, in HE decision modeling the term *uncertainty* is frequently used instead of variance. Uncertainty can also be used in a wide sense to refer to any measure of variability of HE model outcomes. According to Briggs et al. [11], four different types of uncertainty can be distinguished in HE modeling. Although all these types of uncertainty are important, in this article we focus on parameter uncertainty (the uncertainty in model outcomes resulting from uncertainty in the estimation process of the input parameters of an HE model). In HE models, parameter uncertainty is represented in different ways, usually by an uncertainty range containing the predicted model point estimate, a cost-effectiveness plane showing the results of a probabilistic sensitivity analysis (PSA), or a cost-effectiveness acceptability curve [12,13]. Although value of information methods are widely applied to determine whether the uncertainty in HE model outcomes is acceptable for proper decision making [14], model bias can be assessed only by comparing HE model outcomes with empirical data.

The aim of this article was twofold. First, the applicability of several existing validation techniques for comparing HE model outcomes with empirical data is discussed, with special focus on statistical testing. After that, a new method for operational validation is proposed which is aimed at establishing how well HE model outcomes compare with empirical data. In this new method, a level of accuracy that the HE model outcomes should meet is set in advance. The basic idea behind it is rather straightforward: If the model result falls within the limits determined by the required accuracy, then the model result is considered valid. The proportion of valid results obtained in a PSA defines a quantitative measure of the validity of the HE model. Our method is embodied in a Bayesian framework, which allows defining such a validity measure as a probability distribution.

Methods

Using the Statistical Methods Right

Researchers should define statistical metrics to assess consistency of HE model results and empirical data [3,5]. Although there is no "gold standard" criterion [15–24], according to the systematic review by Goldhaber-Fiebert et al. [25], the most frequently used metrics of consistency in HE are the relative or absolute difference in model and study point estimates, the overlap of model outcomes with study uncertainty ranges, and formal statistical tests.

Testing the hypothesis of equal means in HE model outcomes and empirical data is often done by calculating a confidence interval for the difference in means. It is also common to present two confidence intervals separately and check whether they overlap. This should be equivalent to a hypothesis testing about equal means. If the required significance level is 5%, then 95% confidence intervals should *not* be used, but 83% to 84% confidence intervals should [26,27]. Quite often HE articles do present 95% confidence intervals, applying too wide intervals to formally test the hypothesis of equal means at a 5% significance level.

A third approach is to check whether the model's expected value is within the confidence interval on the basis of empirical data [23,28] or whether a single empirical point estimate is within the model uncertainty range [29,30]. Presently, no guidance is given on which of these two approaches should be chosen and in which circumstances.

Using the Right Statistical Methods

Law and McComas [31] raise a more philosophical question. Given the fact that a model is always an approximation to the real system, testing the hypothesis whether model and system are the same would automatically result in rejecting the null hypothesis. Therefore, they question whether hypothesis testing is in fact the appropriate statistical approach.

It should also be emphasized that the terms study "confidence intervals" and model "uncertainty ranges" are used [25,30]. This was done to reflect that HE models rarely can result in confidence intervals in the frequentist sense. This raises the question whether we can formally compare these two types of intervals.

Finally, operational validation is defined as a way to determine that the model output has the "accuracy required for the model intended purpose over the domain of its intended applicability" [32–34]. Therefore, the accuracy required from the model will "depend on its intended use and the utility function of the decision-maker" [31]. Thus, it is the task of the decision maker to establish in advance a level of accuracy for the study and model outcomes to be compared, so that the model results can be regarded as valid.

Formalizing Concepts: Defining a Quantitative Measure for Operational Validity

To structure the subsequent discussion, it is helpful to further formalize issues. This section represents an example for illustrative purposes; other options, including the use of nonparametric approaches, are also possible.

Suppose we have a study sample of 100 patients representative for the study population at hand, where X_1, …, X_{100} denote the observations of a certain outcome (e.g., hospital length of stay in weeks). The mean, SD, and standard error (SE) of the outcome can be estimated from these 100 observations (e.g., $\bar{X}$=0.479, SD_X=0.149, SE_X=0.0149), and thus a 95% confidence interval for the mean is CI_X = 0.450 to 0.508. A cohort HE model, for which a PSA has been run to address parametric uncertainty, results in say 250 means ($\bar{Y}_1$,$\bar{Y}_2$,...,$\bar{Y}_{250}$) and an estimate of the sample mean that is obtained from these 250 replications (e.g., $\bar{Y}$=0.478). The SD obtained from the 250 replications is in fact the SE of the mean ($SE_{\bar{Y}}$) and a 95% uncertainty range for the mean of the outcome, usually given by the simulated 2.5% and 97.5% percentiles, is 0.446 to 0.505. Although the empirical confidence interval and the model uncertainty range are usually compared for overlap, it is important to emphasize that comparing (as in a formal *t* test) observed outcomes X_1, …, X_{100} with means of the PSA samples $\bar{Y}_1$,$\bar{Y}_2$,...,$\bar{Y}_{250}$ is technically incorrect. The 100 values for *X* represent individual observations, whereas the values for *Y* represent 250 means that might have been obtained if input values had been slightly different. An extended version of this section can be found in Appendix A, in which the concepts introduced here are discussed for both cohort and patient-level models.

A better way to compare such empirical data with HE model PSA results is as follows. Because the empirical confidence interval has, say, a 0.95 probability of containing the true value, for cohort HE models, we can set this confidence interval as a

"target" and count how many times, out of the 250, the simulated point estimate $\bar{Y}$ is within CI_X. The number of times (or proportion) that the simulated value is within the target confidence interval would give a quantitative measure of the validity of the model outcome. This notion is the basis for our new method presented in the next section.

An Alternative Bayesian Method for Validating HE Model Outcomes

Suppose that empirical data could be collected from $k > 1$ patients. The outcome of interest is denoted by X_1, X_2, …, X_k. The average over k patients is then $\bar{X}=\frac{1}{k}\sum_{j=1}^{k}X_j$. Suppose also that, on the basis of the same empirical data, an interval containing this average and reflecting the required level of accuracy for the HE model results can be set, for example, by the person evaluating the validation status of the model (such as a decision maker). Such an interval will be referred to as *accuracy interval*, and will be denoted by AI_X. When empirical data are collected from a clinical trial, the confidence interval for the empirical data is a reasonable choice for the accuracy interval. Nevertheless, such a confidence interval may not always be available, for example, when input data are derived from the combination of several published sources that did not report empirical confidence intervals. In that situation, an alternative accuracy interval has to be provided. A simple "what if" situation allowing a certain deviation (e.g., 1%, 5%, or 10%) from the empirical average $\bar{X}$ could be applied.

To predict the outcome of interest we assume that an HE cohort model is used (or alternatively that the results from a patient-level model have been aggregated). Furthermore, we assume that the model outcome was obtained n times ($n > 1$) from a PSA, so that $\bar{Y}_1,\bar{Y}_2,\ldots\bar{Y}_n$ denote the n simulated mean values for our outcome of interest. The validation rule proposed is that a decision maker would regard a model result as valid only if a realization of the model outcome $\bar{y}_i(i=1,\ldots,n)$ is in the interval AI_X. We will denote this as $A_i=I_{\{\bar{y}_i\in AI_X\}}$, where I denotes the indicator function, so that $A_i=1$ if $\bar{y}_i\in AI_X$ and $A_i=0$ otherwise. Assuming that a realization of A_i can be considered as the result of a Bernoulli trial, we can write:

$$P[A_i=a]=p^a(1-p)^{1-a}, a=0,1,$$

where p is the probability that the HE model outcome will be regarded as valid by the decision maker. We assume then that p is a random variable following a prior beta distribution with parameters α and β. If we assume that the result of a full PSA can be regarded as a binomial process of size n where we will observe s successes (s times the model result will be considered valid) and $n - s$ failures, the posterior distribution of p is then also beta but with updated parameters $\alpha + s$ and $\beta + n - s$ [35].

Case Study

Our method is demonstrated with the help of a case study based on a published diabetes model (Modelling Integrated Care for Diabetes based on Observational data) [29]. MICADO is a dynamic population model following overlapping cohorts of diabetic patients aging over time. Incidence and prevalence of diabetes-related complications and mortality are estimated from Dutch registries and systematic literature reviews. A complete description of the model can be found in the study by Van der Heijden et al. [29]. In this case study, we validate the outcome "number of patients with diabetes who are on dialysis or with end-stage renal disease (ESRD)."

Results

New Method in Practice

The example presented in Table 1 illustrates how the method could be used in practice. The results from the PSA iterations 1, 2, 3, 6, 8, and 9 would be regarded as valid. With this information, the prior probability that the model outcome will be regarded as valid by the decision maker is updated. In the example from Table 1, the prior beta distribution is set at parameters $\alpha = 1$ and $\beta = 1$, which means that our previous belief is that the probability that the model outcome is valid for the decision maker is 0.5 with high uncertainty, represented by the 95% credible interval 0.025 to 0.975). After a PSA is run, the probability that the model outcome is valid is updated to 0.58 and the uncertainty is reduced because the posterior 95% credible interval is 0.308 to 0.833), which is much narrower than the previous one.

Case Study

Simulated data were obtained as the results of 300 PSA runs in MICADO. Because MICADO is a cohort model, each PSA outcome can be interpreted as a mean over an unspecified number of simulated individual patients. Empirical data were obtained from the countrywide registration of diabetic patients with ESRD per year [36]. Figure 1 shows the histogram of the 300 PSA outcomes simulated in MICADO. The vertical dashed line is the number of new diabetic patients with ESRD per year from the empirical data (277) and the vertical dotted line is the number of new diabetic patients with ESRD per year predicted by the model (245). The difference between these two represents the prediction error due to bias. The width of the histogram represents a measure of the prediction error due to parameter uncertainty.

The question now is how much deviation from the observed value are decision makers willing to accept as a valid result. To assess this question we consider several possible accuracy intervals that are defined by considering a certain percentage of deviation from the observed number of patients. We have chosen these symmetric accuracy intervals for simplicity but other options are also possible. Note, however, that for this particular outcome (the number of new diabetic patients with ESRD per year), the lower bound of the accuracy interval must always be non-negative, whereas for the upper bound there is in principle no prespecified upper limit (e.g., it can go beyond a 100% deviation). The limits of some of these accuracy intervals

Table 1 – Example of the new method with 10 PSA outcomes.

Prior beta ($\alpha = 1, \beta = 1$)	PSA replication ($A_i=1$ if $\bar{y}_i\in AI_X$ and $A_i=0$ otherwise)										Posterior beta ($\alpha' = 7, \beta' = 5$)
$P[A=1] = \alpha/(\alpha+\beta) = 0.5$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$	$\bar{y}_6$	$\bar{y}_7$	$\bar{y}_8$	$\bar{y}_9$	$\bar{y}_{10}$	$P[A=1 \mid S=6] = \alpha'/(\alpha'+\beta') = 0.58$
95% CI = 0.025–0.975	1	1	1	0	0	1	0	1	1	0	95% CI = 0.308–0.833

Notes. $A = 1$ when the outcome is considered valid.
AI_X, accuracy interval; CI, credible interval; PSA, probabilistic sensitivity analysis; S, number of valid PSA outcomes; $\bar{y}_i$, PSA outcome.

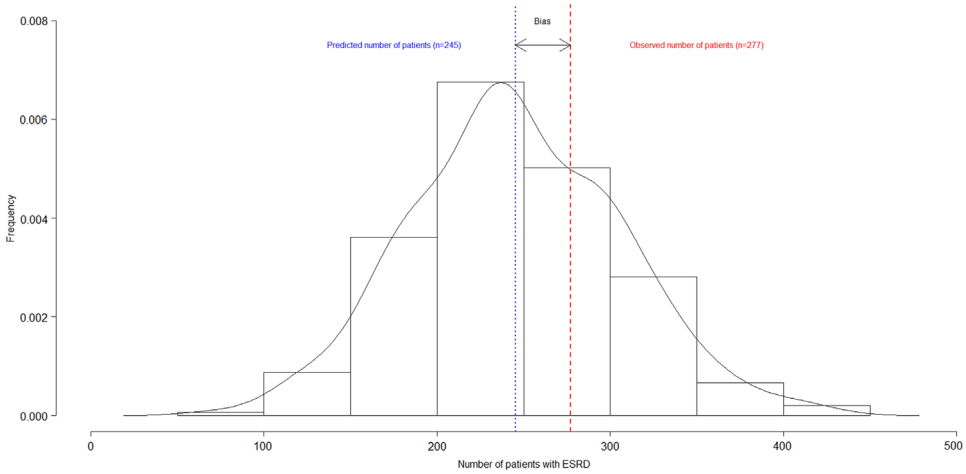

Fig. 1 – Histogram for simulated number of patients with ESRD. ESRD, end-stage renal disease.

are presented in Table 2. Note that only for a relatively large accuracy interval (25% deviation from the observed number of patients given in Table 2) the expected validity is more than 60%, which implies that the model may be considered valid only given relatively low requirements on accuracy for this specific outcome. It is clear that when all PSA outcomes fall within a certain accuracy interval the procedure should be stopped. In Table 2 we can observe that this occurs when the percentage of deviation is 75%. Widening the accuracy interval does not change the expected posterior validity, which in this case will be at most 0.997 (it will never reach 1 because $\beta = 1$ from the prior distribution). This will also set the upper limit for the *x*-axis in Figure 2, in which the posterior probability that the model outcome is considered valid (with 95% credible intervals) has been plotted assuming different accuracy intervals ranging from 0% to 80% deviation from the observed value (we have chosen 80% to show that beyond 75% nothing changes). This confirms that indeed a high probability of a valid outcome is associated with relatively wide accuracy intervals. Whether this represents a problem for decision making depends on the implications of the current outcome (i.e., "number of patients with diabetes who are on dialysis or with end-stage renal disease [ESRD]") for the model main outcome, usually the incremental cost-effectiveness ratio (ICER). This will, for example, depend on the costs, utility, and life-years lost associated with this outcome.

Discussion

The validation of HE models involving comparison of HE model outcomes against empirical observations is of key importance [4,37]. Actual validation efforts quite often present results in an informal way, involving subjective judgment of graphs or comparison of point estimates. Guidelines for validation of HE models lack specific guidance for the operationalization of statistical validation efforts, whereas various quantitative/statistical metrics to assess consistency of model outcomes and empirical data seem to be in use [15–23]. Confidence intervals or hypothesis tests are the preferred statistical methods to assess validation of HE models. Nevertheless, several arguments exist against applying confidence intervals or hypothesis tests as set out in previous sections [31–34].

The method for operational validation proposed in the present article departs from classical statistical techniques. It aims at establishing how well HE model outcomes compare with empirical data by quantifying the degree of validity statistically. Model accuracy is defined as the combination of bias and parameter uncertainty, and our method is concerned with determining whether the current model bias and parameter uncertainty are acceptable for decision making. Note that when the model is biased, reducing parameter uncertainty will reduce the resulting degree of validity when our method is applied, because less PSA outcomes may fall within the accuracy interval.

In our method, validity is operationalized in a Bayesian way. Because in principle a PSA should provide a large number of observations, the choice of a prior distribution should hardly have any influence on the posterior distribution. If new data become available, an HE model could be validated iteratively, by re-estimating the expected posterior probability that the model result will be regarded as valid by the decision maker. The first consequence of having new data is that the target accuracy interval may change. To properly fit with the idea of validation against independent data (i.e., the data used to validate an HE model preferably should not have been used to obtain the model estimates), in a first step an independent validation against the new data could be performed [15–17,19,38]. In a second step, if the model is deemed invalid, the new data could be incorporated into the HE model. The development of an HE model in such a way can also be regarded as a Bayesian process, in which unknown model parameters have statistical prior distributions. Fitting the model to the new empirical data implies that these prior distributions would be updated to posterior distributions that combine the previous information with the new empirical data, resulting in a refitted model. Because the input parameters would be updated, the PSA should be run again (new likelihood). Note that at this point, only dependent (internal) validation is possible because all the available data have been included in the model. The posterior distribution before the new data were available would be the prior now, and on the basis of the new PSA, a new posterior would be obtained that would be compared against the new accuracy interval. This process will increase the model's validation status. If new empirical data were available, then this process of refitting the model and validating against external data can be repeated until the model is deemed valid.

Decision makers should establish the required accuracy level beforehand, ideally in collaboration with stakeholders [39], because they will judge the validity of the HE model results and will have to use these results in their decisions. How to define an

Table 2 – Validation results for the outcome "number of diabetic patients with ESRD".

Deviation from observed number of patients (%)	AI		α'	β'	Expected (posterior) validity*	Posterior validity 95% CI†
	Lower limit	Upper limit				
1	274.23	279.77	8	293	0.027	0.012–0.047
5	263.15	290.85	36	265	0.120	0.085–0.159
10	249.30	304.70	83	218	0.276	0.227–0.328
25	207.75	346.25	203	98	0.674	0.621–0.726
50	138.50	415.50	287	14	0.953	0.927–0.974
75	69.25	484.75	300‡	1	0.997	0.998–1.000
76	66.48	487.52	300	1	0.997	0.998–1.000
80	55.40	498.60	300	1	0.997	0.998–1.000

AI, accuracy interval; α', number of PSA results within the accuracy interval + 1 (note that "+1" is added because as a prior distribution a beta with parameters $\alpha = 1$ and $\beta = 1$ was chosen, but this is just one possibility); β', number of PSA results outside the accuracy interval + 1; CI, credible interval; ESRD, end-stage renal disease; PSA, probabilistic sensitivity analysis.

* Calculated as the expected value of a beta distribution with parameters α' and β'.

† Calculated as the 2.5% and 97.5% percentiles of a beta distribution with parameters α' and β'.

‡ The PSA data from the model are the results of 300 runs. The first run, however, is a deterministic run in which all parameters were set to their mean estimate. Therefore, it was not included here.

accuracy interval might be hard in practice. As a good starting point, the confidence interval of the empirical data, provided that it exists, could be used as guidance for decision makers, because a confidence interval is the most common form of presenting variability for empirical data, and this reflects common practice, for example, in prediction modeling. Nevertheless, a range of accuracy intervals, as shown in our case study, could also be defined when the empirical data do not directly point at a certain interval or when the interval is too wide to be informative. Note also that decision-maker requirements on validation do not need to be based on empirical studies. In this article, empirical data have been chosen for defining accuracy interval as an illustrative example, because it is probably the most straightforward one. Our method, however, can still be used for other types of accuracy intervals, for instance, on the basis of the result of indirect comparisons or network meta-analyses. Our approach results in a posterior distribution around the probability that the HE model outcome will be regarded as valid by the decision maker. This can be reported as an expected value with a credible interval, or graphically plotting this posterior probability for different accuracy intervals, as shown in Figure 2. This figure resembles a cost-effectiveness acceptability curve [12,13], but it reflects the probability that the model is considered valid for this specific outcome.

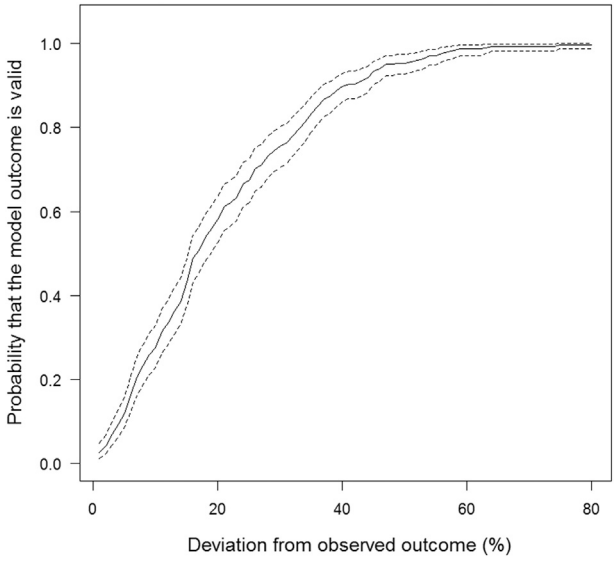

Fig. 2 – Model outcome validity curve—number of patients with ESRD. ESRD, end-stage renal disease.

Independent validation may raise some practical problems because it is common practice to build HE models on the basis of all the best evidence available [40]. Thus, there may be no data to validate the HE model independently. In that case, the validation is called dependent. With sufficient data available, cross-validation techniques exist to keep some of the data for validation purposes. In practice, different parameters of HE models are often estimated on the basis of different sets of empirical data or literature sources. As a result, cross-validation techniques may be less applicable, because these mainly work when a single data set is used as the main source for all model parameters.

In our case study, we have validated the outcome "number of diabetic patients with ESRD." In HE models, this type of outcomes is referred as intermediate, as opposed to a final or main model outcome, which is usually reported in the form of an ICER. HE models should calculate and report enough intermediate outcomes to ensure that the validation process is useful. It is important to emphasize that each of these intermediate outcomes may influence the ICER in a different way. Therefore, the accuracy required for the different intermediate outcomes is not likely to be the same for all of them. Depending on how sensitive the ICER is to changes on each intermediate outcome, it may be reasonable to ask for more or less accuracy for some intermediate outcomes. For example, in our case study, if the outcome "number of diabetic patients with ESRD" had a small impact on the ICER, low accuracy can be required. In that case, the accuracy interval could be wide (e.g., 25% or 50% deviation from the observed number of patients), and given the expected posterior probability in Table 2, a decision maker should most likely consider the model outcome valid. In contrast, with a large impact on the ICER, and a higher required accuracy, the model outcome would be considered invalid.

In this article, we have focused on just one model outcome. In general there would be more than one, and if we aim to give an overall measure of validation, all the outcomes should be compared simultaneously. In a classical frequentist setting, this would require simultaneous hypothesis testing, which can be difficult in general because many assumptions (normality, independence, etc.) have to be checked. In our setting, checking these assumptions is not an issue. Nevertheless, because accuracy intervals are defined separately for each model outcome, to get a high probability of overall validity may require a very high probability for each outcome, which might be unrealistic to reach in practice and not required if their separate effect on the ICER is considered. How to handle this situation is a topic for further research.

In a cohort model there is often a clear timing over the disease's course. Most of these models start with a cohort of (usually newly diagnosed) patients, for example, of the same age and with the same duration of disease, which are then updated after every model cycle. This is not necessarily the case and cohorts based on real-world data, such as the renal registry used in our case study, can represent mixtures of patients with various disease durations and of different ages. In the model used in our case study, patients represent a cohort of typical Dutch diabetic patients with average disease duration (10 years). In that sense, the model deviates from a typical Markov cohort model, because the starting population is not a cohort of newly diagnosed cases. Outcomes that are aggregated over time should be validated with extra caution because a model might result in an overall valid value while having an invalid distribution over the different years. The latter could potentially have serious implications for costs and other outcomes. Nevertheless, it is important to emphasize that the outcome in our case study represents the total annual number of new ESRD cases for the entire diabetic Dutch population but not the total number over the model's time horizon. In particular, it was considered that the number of patients with incident ESRD in 2003 (from the renal registry) could be compared with the model outcome, because both reflected the number of new ESRD cases in a year for the total diabetic Dutch population in 2003. Thus, time dependency is not an issue for the outcome chosen in our case study (because it was generated using a model time horizon of only 1 year). To check the validity of time-dependent outcomes, observations from empirical data over time (since the disease start and over the disease progression pathway) are needed. Given that information, in principle our method could be applied to identify at which time points the model outcome is deemed valid. How to define the overall validity of the outcome over the entire time horizon can present similar difficulties to those discussed when assessing multiple outcomes (e.g., it might be defined as being valid at all time points, but accuracy requirements might be different in time too). Therefore, this is also considered a subject for further research. Nevertheless, as shown earlier, careful selection of outcomes and matching of empirical data can partly avoid the problem.

Conclusions

Current practice in HE decision modeling lacks a consistent standard of reporting on comparisons of model outcomes with empirical data. Existing methods are diverse and not always applied or interpreted correctly. It helps to structure these, by paying close attention to variability in empirical data and uncertainty in model outcomes, and the correct interpretation of confidence intervals. Current practice in HE model validation can be improved by using (in combination with existing correct approaches to assessing validation performance) an alternative method based on assessing whether the model outcomes reach a predefined level of accuracy. The new method presented in this article can be used to validate HE models when the parameter uncertainty is assessed via PSA. Because PSA is the preferred method to study parameter uncertainty in the HE literature, this new method can be applied to the vast majority of HE models. Furthermore, this new method assesses both model bias and parameter uncertainty; thus, the amount of uncertainty predicted by the stochastic model is being assessed. It results in a quantitative measure of the degree of operational validity (expected value and credible interval), where a model that predicts the mean of an outcome correctly but with an overly wide credible interval will be regarded as less valid than a model that predicts the mean correctly with high certainty. This is an advantage over other methods, in which a wide interval may even contribute to a better result of model validity (such as methods that assess whether a target value falls within a credible interval).

Source of financial support: This study was funded by the Netherlands Organization for Health Research and Development (grant no. 80-82500-98-12211).

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2017.04.016.or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Franken M, Nilsson F, Sandmann F, et al. Unravelling drug reimbursement outcomes: a comparative study of the role of pharmacoeconomic evidence in Dutch and Swedish reimbursement decision making. Pharmacoeconomics 2013;31:781–97.
[2] Hoomans T, Severens J, Roer N, Delwel G. Methodological quality of economic evaluations of new pharmaceuticals in the Netherlands. Pharmacoeconomics 2012;30:219–27.
[3] ISPOR-AMCP-NPC Modeling CER Task Forces. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health 2014;17:174–82.
[4] Vemer P, Krabbe PFM, Feenstra TL, et al. Improving model validation in health technology assessment: comments on guidelines of the ISPOR-SMDM Modeling Good Research Practices Task Force. Value Health 2013;16:1106–7.
[5] Eddy DM, Hollingworth W, Caro JJ, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Med Decis Making 2012;32:733.
[6] Hammerschmidt T, Goertz A, Wagenpfeil S, et al. Validation of health economic models: the example of EVITA. Value Health 2003;6:551–9.
[7] Hoerger TJ, Wittenborn JS, Segel JE, et al. A health policy model of CKD: 1. model construction, assumptions, and validation of health consequences. Am J Kidney Dis 2010;55:452–62.
[8] Sendi PP, Craig BA, Pfluger D, et al. Systematic validation of disease models for pharmacoeconomic evaluations. J Eval Clin Pract 1999;5:283–95.
[9] Siebert U, Sroczynski G, Hillemanns P, et al. The German Cervical Cancer Screening Model: development and validation of a decision-analytic model for cervical cancer screening in Germany. Eur J Public Health 2006;16:185–92.
[10] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY: Springer, 2001.
[11] Briggs AH, Weinstein MC, Fenwick EAL, et al. ISPOR-SMDM Modeling Good Research Practices Task Force. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. Value Health 2012;15:835–42.
[12] Briggs AH. Handling uncertainty in cost-effectiveness models. Pharmacoeconomics 2000;17:479–500.
[13] Van Hout BA, Al MJ, Gordon GS, Rutten FFH. Costs, effects and C/E-ratios alongside a clinical trial. Health Econ 1994;3:309–19.
[14] Corro Ramos I, Rutten-van Mölken MPMH, Al MJ. The role of value-of-information analysis in a health care research priority setting: a theoretical case study. Med Decis Making 2013;33:472–89.
[15] Dini FL, Ballo P, Badano L, et al. Validation of an echo-Doppler decision model to predict left ventricular filling pressure in patients with heart

failure independently of ejection fraction. Eur J Echocardiogr 2010;11:703–10.
[16] Kalogeropoulos A, Psaty BM, Vasan RS, et al. Cardiovascular Health Study. Validation of the health ABC heart failure model for incident heart failure risk prediction: the Cardiovascular Health Study. Circ Heart Fail 2010;3:495–502.
[17] Kim LG, Thompson SG. Uncertainty and validation of health economic decision models. Health Econ 2010;19:43–55.
[18] McEwan P, Foos V, Palmer JL, et al. Validation of the IMS CORE Diabetes Model. Value Health 2014;17:714–24.
[19] Pagano E, Gray A, Rosato R, et al. Prediction of mortality and macrovascular complications in type 2 diabetes: validation of the UKPDS outcomes model in the Casale Monferrato Survey, Italy. Diabetologia 2013;56:1726–34.
[20] Palmer AJ, Roze S, Valentine WJ, et al. Computer modeling of diabetes and its complications: a report on the Fourth Mount Hood Challenge Meeting. Diabetes Care 2007;30:1638–46.
[21] Palmer AJ, Roze S, Valentine WJ, et al. Validation of the CORE Diabetes Model against epidemiological and clinical studies. Curr Med Res Opin 2004;20(Suppl. 1):S27–40.
[22] Palmer AJ, The Mount Hood 5 Modeling Group. Computer modeling of diabetes and its complications: a report on the Fifth Mount Hood Challenge Meeting. Value Health 2013;16:670–85.
[23] Perreault S, Levinton C, Laurier C, et al. Validation of a decision model for preventive pharmacological strategies in postmenopausal women. Eur J Epidemiol 2005;20:89–101.
[24] Willis M, Johansen P, Nilsson A, et al. Validation of the Economic and Health Outcomes Model of Type 2 Diabetes Mellitus (ECHO-T2DM). Pharmacoeconomics 2017;35:375.
[25] Goldhaber-Fiebert JD, Stout NK, Goldie SJ. Empirically evaluating decision-analytic models. Value Health 2010;13:667–74.
[26] MacGregor-Fors I, Payton ME. Contrasting diversity values: statistical inferences based on overlapping confidence intervals. PLoS One 2013;8: e56794.
[27] Payton ME, Greenstone MH, Schenker N. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? J Insect Sci 2003;3:34.
[28] Wong J, Chase JG, Hann CE, et al. A subcutaneous insulin pharmacokinetic model for computer simulation in a diabetes decision support role: validation and simulation. J Diabetes Sci Technol 2008;2:672–80.
[29] Van der Heijden AA, Feenstra TL, Hoogenveen RT, et al. Policy evaluation in diabetes prevention and treatment using a population-based macro simulation model: the MICADO model. Diabet Med 2015;32:1580–7.
[30] Leunis A, Redekop WK, van Montfort KAGM, et al. The development and validation of a decision-analytic model representing the full disease course of acute myeloid leukemia. Pharmacoeconomics 2013;31:605–21.
[31] Law AM, McComas MG. How to build valid and credible simulation models. In: Peters BA, Smith JS, Medeiros DJ, Rohrer MW, eds., Proceedings of the 2001 Winter Simulation Conference, IEEE, New York, 2001;22-9.
[32] Sargent RG. Verification, validation, and accreditation of simulation models. In: Joines JA, Barton RR, Kang K, Fishwick PA, eds., Proceedings of the 2000 Winter Simulation Conference, IEEE, New York, 2000;50–9.
[33] Sargent RG. Some approaches and paradigms for verifying and validating simulation models. In: Peters BA, Smith JS, Medeiros DJ, Rohrer MW, eds., Proceedings of the 2001 Winter Simulation Conference, IEEE, New York, 2001;106–14.
[34] Sargent RG. Validation and verification of simulation models. In: Ingalls RG, Rossetti MD, Smith JS, Peters BA, eds., Proceedings of the 2004 Winter Simulation Conference, IEEE, New York, 2004;17–28.
[35] Lee PM. Bayesian Statistics: An Introduction. Hoboken, NJ: Wiley, 2012.
[36] Oppe M, Barendregt W, Treur MJ. Statistical Report 2007. The Netherlands Renal Registry RENINE, 2007.
[37] Vemer P, Corro Ramos I, van Voorn GAK, et al. AdViSHE: a validation-assessment tool of health-economic models for decision makers and model users. Pharmacoeconomics 2016;34:349–61.
[38] Eddy DM. The frequency of cervical cancer screening: comparison of a mathematical model with empirical data. Cancer 1987;60:1117–22.
[39] Van Voorn GAK, Vemer P, Hamerlijnck D, et al. The Missing Stakeholder Group: why patients should be involved in health economic modelling. Appl Health Econ Health Policy 2016;14:129–33.
[40] Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice of decision analytic modeling in health care evaluation: report of the ISPOR Task Force on Good Research Practices—modeling studies. Value Health 2003;6:9–17.