PharmacoEconomics (2017) 35:613–624
DOI 10.1007/s40273-017-0494-4

PRACTICAL APPLICATION

Bayesian Methods for Calibrating Health Policy Models: A Tutorial

Nicolas A. Menzies[1,2] · Djøra I. Soeteman[2] · Ankur Pandya[2,3] · Jane J. Kim[2,3]

Published online: 28 February 2017

Abstract Mathematical simulation models are commonly used to inform health policy decisions. These health policy models represent the social and biological mechanisms that determine health and economic outcomes, combine multiple sources of evidence about how policy alternatives will impact those outcomes, and synthesize outcomes into summary measures salient for the policy decision. Calibrating these health policy models to fit empirical data can provide face validity and improve the quality of model predictions. Bayesian methods provide powerful tools for model calibration. These methods summarize information relevant to a particular policy decision into (1) prior distributions for model parameters, (2) structural assumptions of the model, and (3) a likelihood function created from the calibration data, combining these different sources of evidence via Bayes' theorem. This article provides a tutorial on Bayesian approaches for model calibration, describing the theoretical basis for Bayesian calibration approaches as well as pragmatic considerations that arise in the tasks of creating calibration targets, estimating the posterior distribution, and obtaining results to inform the policy decision. These considerations, as well as the specific steps for implementing the calibration, are described in the context of an extended worked example about the policy choice to provide (or not provide) treatment for a hypothetical infectious disease. Given the many simplifications and subjective decisions required to create prior distributions, model structure, and likelihood, calibration should be considered an exercise in creating a reasonable model that produces valid evidence for policy, rather than as a technique for identifying a unique theoretically optimal summary of the evidence.

Electronic supplementary material The online version of this article (doi:10.1007/s40273-017-0494-4) contains supplementary material, which is available to authorized users.

✉ Nicolas A. Menzies
nmenzies@hsph.harvard.edu

[1] Department of Global Health and Population, Harvard T.H. Chan School of Public Health, 665 Huntington Ave, Boston, MA 02115, USA

[2] Center for Health Decision Science, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

[3] Department of Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

Key points for decision makers

Calibration describes the process of estimating the parameters of a simulation model so that the model predictions are consistent with external data that may be available.

Calibration can ensure that the simulation model provides a realistic depiction of the processes determining the outcomes of a health policy decision, and thereby provide more accurate predictions for outcomes of interest (e.g., summary measures of health burden, budget impact estimates, cost-effectiveness ratios).

Bayesian methods use simple laws of probability to synthesize available evidence on model parameters and modeled outcomes. This tutorial describes the theoretical underpinnings of this approach, provides an extended worked example to show how the approach can be implemented, and discusses practical considerations.

1 Introduction

Mathematical simulation models are commonly employed to assess health policy decisions because empirical studies would be costly or impractical to undertake, or because waiting for empirical results to become available would delay an urgent decision. These models typically combine multiple sources of evidence about how policy alternatives influence outcomes, with this evidence encoded in model parameters and formulae. Some evidence will pertain directly to individual model parameters. For example, a parameter describing sensitivity of a diagnostic test may be directly informed by meta-analyses on the subject, and in this situation, it is generally straightforward to incorporate the evidence into the analysis. Other evidence may describe modeled outcomes, which are typically complex functions of multiple parameters. For example, programmatic evidence on the number of individuals initiating treatment for a particular condition may be related to parameters describing the size of the target population, screening rates, sensitivity and specificity of various tests in the diagnostic algorithm, and potentially other parameters. In this situation, individual parameters cannot be updated directly, and other methods must be used. Model calibration represents the process of incorporating evidence on modeled outcomes into the model.

The ultimate goal of these health policy models is usually to estimate summary outcomes—health burden, budget impact, or cost effectiveness—for which no direct evidence is available, but that are increasingly requested by decision makers. By making best use of all available evidence, in particular, calibrating models to relevant empirical data, we can obtain better estimates for these outcomes of interest. In this tutorial, we provide an introduction to Bayesian calibration approaches and apply them in the context of a worked example. Bayesian approaches use simple probability rules to combine three sources of information: (1) evidence about the distribution of model parameters, (2) evidence about the distribution of modeled outcomes, and (3) model structural assumptions that relate parameters and modeled outcomes. By acknowledging uncertainty in both parameters and calibration data, these methods appropriately weight different sources of evidence and provide estimates of uncertainty in model results.

2 Theoretical Framework

2.1 Probability Model and Bayes' Theorem

Bayesian statistics assumes that evidence about quantities of interest can be represented by probability distributions, with Bayes' theorem providing the machinery for updating a probability distribution with new data. If $p(\theta)$ represents the probability distribution for a quantity θ before considering new data (the prior), and $p(\theta|Y)$ represents the probability distribution for θ incorporating new data Y (the posterior), we can obtain $p(\theta|Y)$ by multiplying the prior by $p(Y|\theta)$ (the likelihood function) and scaling by $p(Y)$, the probability of observing the data:

$$p(\theta|Y) = \frac{p(\theta) \times p(Y|\theta)}{p(Y)}. \tag{1}$$

For most applications, it is sufficient to represent the posterior as proportional to the prior times the likelihood, omitting $p(Y)$:

$$p(\theta|Y) \propto p(\theta) \times p(Y|\theta). \tag{2}$$

By keeping Y fixed and varyingθ, the likelihood function $p(Y|\theta)$ can be used to describe the relative likelihood of different values of θ given the evidence represented by Y. Parameter sets with higher values of $p(Y|\theta)$ are more consistent with the data Y, and this property allows us to assess the extent to which the evidence supports one parameter set compared with another. Bayesian methods focus on estimating the posterior distribution $p(\theta|Y)$, and therefore incorporate evidence from both prior and likelihood. Estimating the posterior distribution can be complicated when calibrating health policy models, but the basic components are those described above: (1) prior distributions representing evidence on model parameters, (2) a likelihood function relating modeled outcomes to empirical data, and (3) the model itself, which translates model parameters into modeled outcomes.

2.2 Application to Health Policy Models

The health policy model (M) is a mathematical function that transforms a parameter set θ into a set of modeled outputs φ. Both θ and φ can be multi-dimensional, θ is a parameter set including a value for each model parameter, while φ can include a variety of outcomes (essentially, all outcomes from the model). Information about θ is operationalized as the prior $p(\theta)$. Evidence on model outcomes is operationalized by the likelihood function $p(Y|\varphi_c)$ [subscript c denoting the subset of outcomes used for calibration].

Bayes' theorem allows us to calculate the posterior for θ:

$$p(\theta|Y) \propto p(\theta) \times p(Y|M(\theta)). \tag{3}$$

Most modern Bayesian analyses approximate this posterior distribution using numerical methods. Rather than producing a closed-form equation that could be used to generate new parameter sets, these methods produce a large sample of parameter sets where each set represents a draw from the posterior parameter distribution $p(\theta|Y)$. Once we have this sample of calibrated parameter sets, we can use these in a traditional Monte Carlo simulation to estimate a distribution for any outcome of interest:

$$p(\varphi|Y) = p(M(\theta|Y)). \quad (4)$$

2.3 Relationship to Other Approaches

The Bayesian approach outlined above is one of several frameworks described for calibrating simulation models. In contrast to ad hoc, frequentist, or otherwise non-Bayesian calibration frameworks [1–4], the theory underlying Bayesian approaches provides an axiomatic basis for deciding how to quantify evidence, avoiding arbitrary decisions about the relative weight to be placed on different data sources or the use of heuristics to select well-fitting parameter sets. In contexts where priors, likelihood, and model are all correctly specified, the Bayesian approach can provide a theoretically optimal summary of the evidence [5, 6], allowing a decision maker to maximize expected utility when paired with a utility function representing his/her preferences for different outcomes [7]. However, for complicated policy problems, the task of specifying priors, likelihood, and model will involve many choices for which the underlying theory provides only general guidance. Simplifications are invariably made in terms of how the simulation model is constructed, or which data are deemed relevant to the decision problem. In this context, the theoretical guarantees described above will not necessary hold. Despite these challenges, Bayesian approaches have been found to perform well compared with alternative approaches [8], and provide a principled framework for making analytic choices. The section below considers these applied issues in the context of a simplified policy problem.

3 Worked Example

3.1 Policy Question

The example deals with a hypothetical sexually transmitted disease. Infected individuals experience progressively increasing mortality rates. A treatment is available, which reduces mortality but is not curative and is required for life. Currently, individuals with late-stage disease are eligible for treatment, but it is unclear whether individuals with early-stage disease should also receive treatment. The analysis is designed to estimate the cost effectiveness of providing treatment for this group, compared with a status quo restricting treatment to individuals with late-stage disease.

3.2 Study Model

The model is adapted from approaches for modeling human immunodeficiency virus in high-burden settings [9], simplified for ease of exposition for this example. The population is divided into five health states including non-susceptible (N), susceptible (S), early disease (E), late disease (L), and treatment (T). The number of individuals by state and year (t) is given by N_t, S_t, E_t, L_t, and T_t, respectively. Individuals enter the model distributed across the N and S states, and transition between states to allow for infection (S to E), disease progression (E to L), treatment initiation (E and L to T), and death (N, S, E, L and T to D) via background and disease-specific mortality. Figure 1 shows the model schematic and Table 1 describes model parameters. Differential equations describing the model are given in the Technical Appendix.

The force of infection (λ_t) is calculated as $\lambda_t = \rho \frac{E_t+L_t}{S_t+E_t+L_t+T_t}$, where ρ represents the effective contact rate (the rate of infection for a susceptible individual exposed to 100% infected contacts). This formulation assumes a fraction of the population (N) is not sexually active.

A model simulation is initiated 30 years in the past ($t = 0$) when the epidemic is believed to have started, and

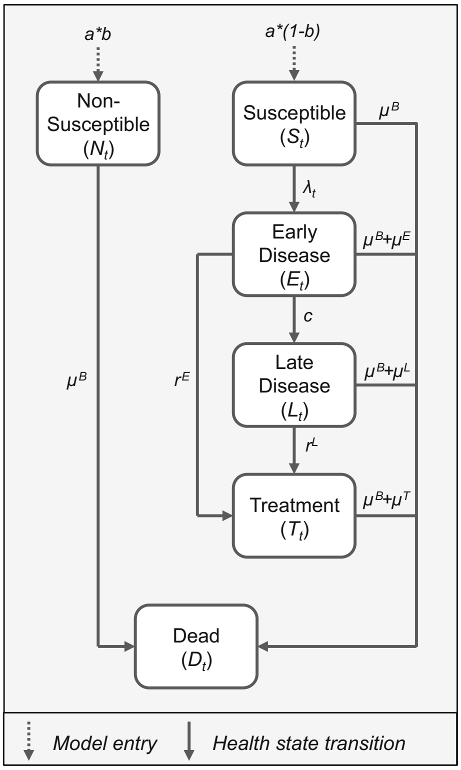


Fig. 1 Schematic of the example model. *Boxes* represent model states and *arrows* represent transitions. Parameter descriptions are provided in Table 1

Table 1 Model parameters and prior distributions

Parameter	Description	Prior distribution	Implied mean and 95% interval
a	Annual birth rate, calculated to achieve steady state based on population size of 1 million and μ^B	No uncertainty, taken as fixed	15,000
b	Fraction of births entering non-susceptible state	Beta, $\alpha = 2$, $\beta = 8$	0.20 [0.03, 0.48]
μ^B	Background mortality rate	No uncertainty, taken as fixed	0.015
μ^E	Disease-specific mortality for early disease	Log-normal, $\mu = -3.121$, $\sigma = 0.5$	0.05 [0.02, 0.12]
μ^L	Disease-specific mortality for late disease	Log-normal, $\mu = -1.511$, $\sigma = 0.5$	0.25 [0.08, 0.59]
μ^T	Disease-specific mortality on treatment	Log-normal, $\mu = -3.814$, $\sigma = 0.5$	0.025 [0.01, 0.06]
ρ	Effective contact rate for transmission	Log-normal, $\mu = -0.818$, $\sigma = 0.5$	0.025 [0.01, 0.06]
p	Rate of progression from early to late disease	Log-normal, $\mu = -2.428$, $\sigma = 0.5$	0.10 [0.03, 0.24]
r^L	Rate of treatment uptake for late disease	Log-normal, $\mu = -0.818$, $\sigma = 0.5$	0.50 [0.17, 1.18]
r^E	Rate of treatment uptake for early disease	No uncertainty, taken as fixed	0.0
c^T	Annual cost of treatment	Log-normal, $\mu = 6.888$, $\sigma = 0.2$	US$1000 [662, 1451]

run forward to provide historical and future estimates for various outcomes. The analysis adopts a 20-year time horizon, with the incremental cost-effectiveness ratio (ICER) representing the ratio of incremental costs to incremental life-years lived for the proposed policy vs. status quo (both undiscounted). Additional details (including R code) are provided in the Technical Appendix.

3.3 Analytic Results Without Calibration Data

We first discuss how an analysis might proceed if calibration data were not available. In this situation, it is common to calculate results using point estimates for each parameter (i.e., ignoring parameter uncertainty), or alternately to calculate results using a probabilistic sensitivity analysis. Using parameter point estimates is the simplest approach, but unlike probabilistic sensitivity analysis, this approach does not provide uncertainty estimates for modeled outcomes. In addition, as modeled outcomes are typically non-linear functions of the parameters, the results produced by using parameter point estimates will differ from the mean results from a probabilistic sensitivity analysis (via Jensen's inequality) [10]. For these reasons, it is becoming increasingly conventional for modeled economic evaluations to take parameter uncertainty into account [11–13]. Probabilistic sensitivity analysis (achieved via Monte Carlo simulation) is the accepted approach for quantifying the implications of parameter uncertainty for modeled outcomes and providing summary information on decision uncertainty [14, 15]. With this approach, important decisions must be made on how to synthesize evidence on parameter values and operationalize this evidence as prior distributions. The process of defining model parameters and creating prior distributions is not discussed here as it has been addressed in detail elsewhere [12], but it is an important consideration during model calibration. The process of identifying and synthesizing evidence for both priors and calibration targets is best viewed as components of a single estimation procedure [16–20].

We conducted a probabilistic sensitivity analysis to demonstrate the results that would be obtained with the example model in the absence of calibration data. Many parameter sets were drawn from the prior distributions described in Table 1, and outcomes estimated by Monte Carlo simulation. From the uncalibrated model, incremental life-years were estimated to be 213,000 (equal-tailed 95% interval = [6, 775]), and incremental costs were estimated to be US$277 million [−34, 1235]. The ICER was estimated to be US$1300 per life-year saved, though this estimate is very imprecise because of parameter uncertainty (Fig. 5).

3.4 Operationalizing Calibration Targets

The calibration data include (1) population-based prevalence surveys for the present year as well as 10 and 20 years ago ($t = 10$, 20, and 30); (2) natural history studies reporting life expectancy of 10 [8, 12] years among newly infected individuals without treatment; and (3) program data estimating current treatment volume at 75,000, ± 5000. Table 2 describes these data and how they can be operationalized as calibration targets, which are mathematical functions that summarize the available evidence on modeled outcomes.

3.4.1 *Whether to Use the Original Data Likelihood*

For the prevalence calibration targets in the worked example, we know the data generation process, a simple random sample of the entire population, using a perfect test. This data generation process produces three independent binomial likelihoods that can be used directly as calibration targets. In other situations, the process that generates calibration data (how data are collected, how sample estimates relate to population characteristics) will be more complex, and may produce a likelihood requiring a complicated series of calculations and/or involve additional parameters to be estimated simultaneously. This can present challenges for using the data likelihood itself as a calibration target, particularly if evaluating the likelihood is computationally intensive. An alternative approach is to analyze the data using conventional methods to produce summary statistics (e.g., point estimates and confidence intervals), and use these to create calibration targets. This approach can reduce the difficulty of creating calibration targets, and allow the appropriate analytic techniques to be applied to the data. However, care must be taken that this two-step approach does not omit important features of the data, such as correlations between variables in the likelihood.

3.4.2 Choosing a Function to Approximate the Likelihood

When creating calibration targets from published research, the underlying data are generally not available and the likelihood must be approximated with a different function. This is the case with the calibration target for average survival, where the likelihood is approximated by a normal distribution parameterized using the published mean and confidence interval. Even when data are non-normal, the central limit theorem dictates that the normal distribution will provide an increasingly accurate approximation of the true likelihood as the sample size increases, and given its analytic tractability it can usually be included among the options being considered. Care must be taken in situations where the outcome being calibrated is restricted to some part of the number line (such as average survival, which is strictly positive). In this situation, if a substantial portion of the normal distribution used as a calibration target lies outside of the support of the outcome, the mean and variance of the calibration target could be biased.

Table 2 Calibration data and likelihoods

Empirical data	Relevant model outcome	Likelihood function
Three population-based surveys of disease prevalence, each a simple random sample of 500 individuals, using a diagnostic algorithm with perfect sensitivity and specificity At $t = 10$: 25 out of 500 positive At $t = 20$: 75 out of 500 positive At $t = 30$: 50 out of 500 positive	Prevalence in year t can be directly estimated from the model as $\mathrm{prev}_t = \frac{E_t+L_t+T_t}{N_t+S_t+E_t+L_t+T_t}$, evaluated in each year in which the survey was conducted	Independent binomial likelihoods for the form $L_1 = \binom{500}{25}\mathrm{prev}_{10}^{25}(1-\mathrm{prev}_{10})^{500-25}$ The log-likelihoods can be written more parsimoniously: $\ln(L_1) \propto 25\mathrm{prev}_{10} + 475(1-\mathrm{prev}_{10})$ $\ln(L_2) \propto 75\mathrm{prev}_{20} + 425(1-\mathrm{prev}_{20})$ $\ln(L_3) \propto 50\mathrm{prev}_{30} + 450(1-\mathrm{prev}_{30})$
A study published from an observational cohort suggesting a mean survival following infection of 10.0 years, with confidence interval [8.0, 12.0]	Mean survival following infection is not directly estimated by the model, but can be calculated from the parameter values as: $\mathrm{surv} = \frac{1}{\mu_B+\mu_E+p} + \frac{p}{\mu_B+\mu_E+p} \times \frac{1}{\mu_B+\mu_L}$	The sampling distribution for the study data is unknown, and is approximated with a normal likelihood: $L_4 = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(10-\mathrm{surv})^2}{2\sigma^2}\right)$ The standard deviation is estimated from the width of the reported confidence interval:$\hat{\sigma} = \frac{12-8}{2\times 1.96}$ The log-likelihood can be written more parsimoniously: $\ln(L_4) \propto -\frac{(10-\mathrm{surv})^2}{2\hat{\sigma}^2}$
Routine reporting suggests 75,000 individuals receiving treatment in the current year, ±5000 because of uncertainties in reporting	Treatment volume in year t is equal to T_t	The sampling distribution for the study data is unknown, and is approximated with a normal likelihood, with $\hat{\sigma} = \frac{5000}{1.96}$: $\ln(L_5) \propto -\frac{(75{,}000-T_{35})^2}{2\hat{\sigma}^2}$

The choice of distribution to approximate a likelihood can be consequential, especially when the various sources of evidence disagree. In this situation, the slope of the log-density in the tails of the distribution will determine the strength of the calibration target. Figure 2 shows the density function, log-density, and the slope of the log-density of four different non-negative distributions, the Log-normal, Gamma, Weibull, and truncated normal distributions, each parameterized with a mean of 1.0 and a variance of 0.5. The distributions display quite different behavior in their tails, despite matching means and variances. For these reasons, it is important to understand how functions used for calibration targets behave under different scenarios.

3.4.3 Calibrating to Outcomes Not Produced by the Model

Another feature of the calibration target for survival is that the outcome being calibrated (mean survival) is calculated outside the model. As constructed, the model does not produce an estimate for survival; however, this outcome can be calculated as a function of the relevant model parameters. The reason for doing this, specifying an additional model or functional relationship for use in the calibration, is to obtain better estimates for parameters used in the main model. While the evidence on survival could be incorporated into the prior, including this evidence in the calibration is convenient, and also ensures that evidence is represented appropriately.

3.4.4 Non-Sampling Error

The calibration target for current treatment volume is operationalized in the same way as the calibration target for survival. However, in this case the uncertainty represents perceived imperfections in reporting rather than sampling uncertainty. Non-sampling biases represent a separate consideration to the sampling uncertainty captured in conventional likelihood functions. Failing to consider non-sampling biases can lead to calibration targets that are overly strong and/or biased (the Technical Appendix provides further discussion of strong calibration targets). If the bias is poorly understood, the solution may be to arbitrarily weaken the calibration target based on some assessment of the magnitude of the bias, as in the treatment volume example. If the source of bias is well understood, a more objective solution may be available. For example, if the diagnostic used in the prevalence likelihood had imperfect sensitivity and specificity, a likelihood could be constructed for test positivity $(\text{prev} \times \text{sensitivity} + (1 - \text{prev}) \times (1 - \text{specificity}))$ rather than true prevalence. Sensitivity and specificity could be incorporated as point estimates, or given their own priors and treated like the other model parameters.

3.5 Calibration Approaches

3.5.1 Point Estimates or Full Posterior Distribution

To produce estimates of uncertainty around modeled results, the calibration will need to produce a distribution

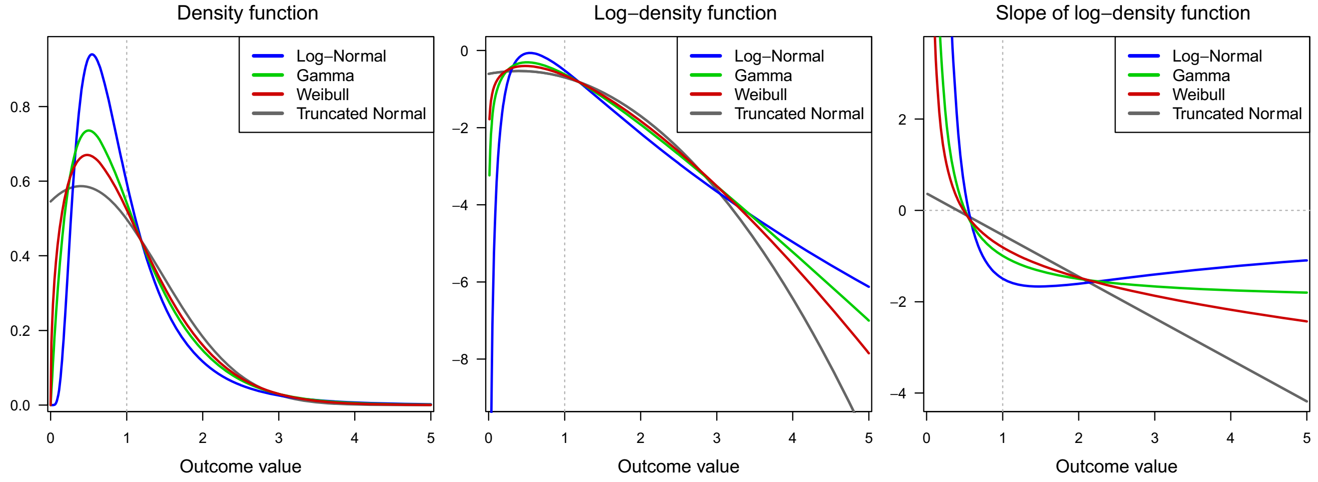


Fig. 2 Density function, Log-density function, and slope of Log-density function for various non-negative distributions with mean = 1.0 and variance = 0.5. At the *lower tail*, the truncated normal will allow values close to zero, which would be heavily penalized by the other distributions. At the *upper tail*, the truncated normal and Weibull will penalize extreme values more strongly than the other two distributions. For the log-normal distribution, the slope of the log-density actually returns towards zero for increasingly extreme values in the upper tail. As a consequence, the strength of the calibration target first increases and then declines for increasing values of the outcome being calibrated. For this reason, the log-normal (and other 'fat-tailed' distributions) can be a problematic choice for creating calibration targets, unless this behavior is specifically desired

of values for model parameters. However, obtaining a single 'best fitting' parameter set can be an important first step. This can be achieved via optimization routines available in most statistical software, identifying the parameter set that maximizes the posterior parameter distribution (i.e., prior distribution times likelihood function). This 'maximum a posteriori' estimate is not necessarily equal to the mean or mode of the modeled outcomes, but will likely be close, and will provide valuable information on how the calibration performs. For example, examining the results may show that fitted values for important parameters are far from the mass of the prior distribution, or the fitted model may not match the calibration targets. While not necessarily problematic, examining these inconsistencies may reveal errors in the way evidence is summarized, problematic assumptions, or simply programming bugs. This process also provides information on how the main calibration will proceed. If the model runs too slowly to obtain a reasonable best-fitting parameter set, this is a signal that steps to improve efficiency may be needed before continuing. If optimization identifies a different 'best-fitting' parameter set from different starting points, this suggests the posterior is multimodal or otherwise difficult to sample. Optimization algorithms can also struggle when some parameters (or functions of parameters) are not identified by the combined evidence in prior and likelihood. Optimization can pose particular challenges for stochastic models, for which the modeled outcomes associated with a particular parameter set are estimated with Monte Carlo error. In this situation, the likelihood for a given parameter set will be estimated with error, and can be computationally expensive to compute. This issue can also cause problems when estimating the posterior distribution. Special issues for stochastic models are discussed in Sect. 3.5.5.

Figure S1 shows fitted values for model parameters and calibration targets obtained via 'maximum a posteriori' estimation. It is good practice to undertake the optimization in log space, that is, to use the optimizer to identify the sum of logged priors plus logged calibration targets, as this can minimize numerical errors. The results shown below were obtained using the Broyden–Fletcher–Goldfarb–Shanno algorithm from R's optim function [21]. In practice, it is useful to try different algorithms to find one that works well for a particular problem.

3.5.2 *Whether to Include All Parameters in the Optimization*

Parameters that are fixed (e.g., a and μ^B in Table 1) can be excluded from the calibration. At a theoretical level, it is unproblematic to include all uncertain parameters in the calibration, but it can be useful to omit some parameters, as often the optimization routine (or algorithm used to sample the posterior distribution) will perform better with a reduced number of parameters. An obvious case where a parameter (or parameters) can be omitted is where the parameters are statistically independent from (1) other model parameters in the prior distribution and (2) the calibration targets. The first condition holds if all prior distributions are independent. The second condition holds if the parameters to be omitted have no mathematical relationship with the calibrated outcomes. This will typically be true for parameters such as unit costs and utility weights. In the worked example, the annual treatment cost (c^T) meets both criteria, and was excluded from the calibration. It may be advantageous to exclude parameters from the calibration in situations where the independence conditions only hold qualitatively, but subjective judgments must be made about whether it is reasonable to do so. For omitted parameters, the 'maximum a posteriori' point estimate is equal to the prior mode, and to obtain random samples one simply samples from the prior.

3.5.3 *Sampling from the Posterior Distribution*

Bayesian statistics is a growing field, and methods for sampling from the posterior parameter distribution are the subject of active development. The first method described below, an application of sampling importance resampling (SIR) [22], is straightforward to implement and explain, and aspects of this algorithm will be familiar to those with knowledge of probabilistic sensitivity analysis (steps 1 and 2 are identical). This algorithm can provide reasonable results when calibration targets are not overly strong relative to the prior. The steps of the algorithm are as follows:

Step 1 Draw a large number of parameter sets from the prior distribution.

Step 2 For each parameter set (θ_i), run the model and estimate modeled outcomes.

Step 3 Using these modeled outcomes, estimate the likelihood for the parameter set, $L(\theta_i)$, and retain this value.

Step 4 Resample from the original parameter sample with replacement, using the likelihood values as sampling weights.

To implement SIR with the worked example, a sample size of 100,000 was used for the initial sample as well as the resample. The resample included 797 unique parameter sets. The reduction in the number of unique parameter sets (three orders of magnitude) occurs because some parameter sets have much higher likelihoods relative to others, and are sampled many times over in step 4. The effective sample size (ESS) is a useful metric for understanding the information value of a weighted sample, describing the size of a simple random sample that would produce a mean with equivalent variance to the weighted sample [23]. ESS is calculated as the squared sum of the sampling weights divided by the summed squares of these weights:

$$\mathrm{ESS} = \left(\sum\nolimits_i L(\theta_i)\right)^2 \Big/ \sum\nolimits_i L(\theta_i)^2. \tag{5}$$

In this example, ESS = 88, and this low ESS indicates non-trivial Monte Carlo error in the study results. This highlights the major drawback of SIR, that it is inefficient. Efficiency will be worse when the prior is very dispersed compared with the likelihood, or covers a different area of the parameter space.

Modern algorithms improve efficiency by concentrating sampling in regions of the parameter space with higher posterior density. One approach developed specifically for mechanistic models is incremental-mixture importance sampling (IMIS) [24]. Using IMIS, we obtained a posterior sample of 10,000 parameter sets including 6372 unique parameter sets, with ESS = 4713. The algorithm evaluated the model 27,600 times, and using the ratio of ESS to total samples as a measure of efficiency, IMIS was 150–200 times more efficient than SIR. SIR may be adequate when the model can be evaluated easily and when the calibration targets are not very strong, otherwise more efficient techniques will be needed.

Other estimation approaches use Markov chain Monte Carlo methods [25, 26]. Where the simulation model is relatively simple, these methods can be implemented using Bayesian modeling packages such as WinBUGS [27] or Stan [28], where the model is specified in the package's modeling language and the software takes care of the estimation details [18, 19]. In situations where the likelihood is intractable or computationally expensive to estimate (commonly in microsimulation models), the likelihood can be replaced by summary statistics calculated from simulated data, and the posterior approximated using approximate Bayesian computation methods [29, 30].

3.5.4 Evaluating Model Fit

Once a sample of fitted parameter sets has been obtained, it is important to evaluate the model fit (this is in addition to reviewing the convergence diagnostics appropriate to the fitting algorithm used [24, 31, 32]). This step involves reviewing the posterior distribution of model parameters against their priors, and model predictions against calibration targets. Figure 3 compares marginal prior and posterior distributions for the calibrated parameter sets produced by IMIS, showing the extent to which the prior and the posterior overlap. In particular, it can be seen that for parameters ρ and b the distribution of the posterior is substantially narrower than the prior, indicating that these parameters are primarily identified through the calibration targets. Figure 4 compares model predictions with calibration targets, allowing the consistency of the model predictions with the calibration targets to be evaluated graphically. For all of these outcomes, the model predictions closely match the calibration targets. This figure also shows results for several other modeled outcomes for which no calibration target was available, and it is good practice to review all outcomes for which there might be some prior understanding about what appropriate values might be, even if this evidence has not been included in the calibration. In some cases, comparing model predictions with calibration targets can be achieved by using the model to generate simulated data, which can be compared with the data used to construct the calibration likelihood (posterior predictive checks [33, 34]). Diagnostics have also been developed for assessing parameter identifiability in the fitted model [35].

Where discrepancies are observed they should be investigated, and the knowledge this generates may lead to revisions to priors, likelihood, or model. However, iterative tweaking of these components for the sole purpose of maximizing calibration fit should be avoided. For those models able to predict a long list of outcomes, it is unreasonable to expect perfect fit with all available empirical data, as the model is invariably a simplification of a very complicated reality. Pursuing this goal can simply result in overfitting, risking overconfidence in model predictions and biased estimates for the outcomes of interest. Assessing the influence of individual parameters can be difficult in the context of a calibrated model, and for this task, partial rank correlation coefficients can be used [36, 37].

3.5.5 Special Issues for Stochastic Models

Several of the calibration approaches described in the preceding sections become more difficult to apply in the context of stochastic models, such as stochastic microsimulations or agent-based models. One issue is computing time, as the methods described here require the model to be evaluated a large number of times, and stochastic models can be computationally expensive to evaluate. Another issue arises from the fact that the results from these models are estimated with Monte Carlo error. This issue will likely pose the greatest problems in the context of optimization, as Monte Carlo error around modeled outcomes will produce random error in the likelihood estimated for a given parameter set. In this situation, gradient-based optimization methods may struggle. An optimization approach that is robust to stochastic uncertainty (such as simulated annealing) may be needed, though these approaches can be less efficient.

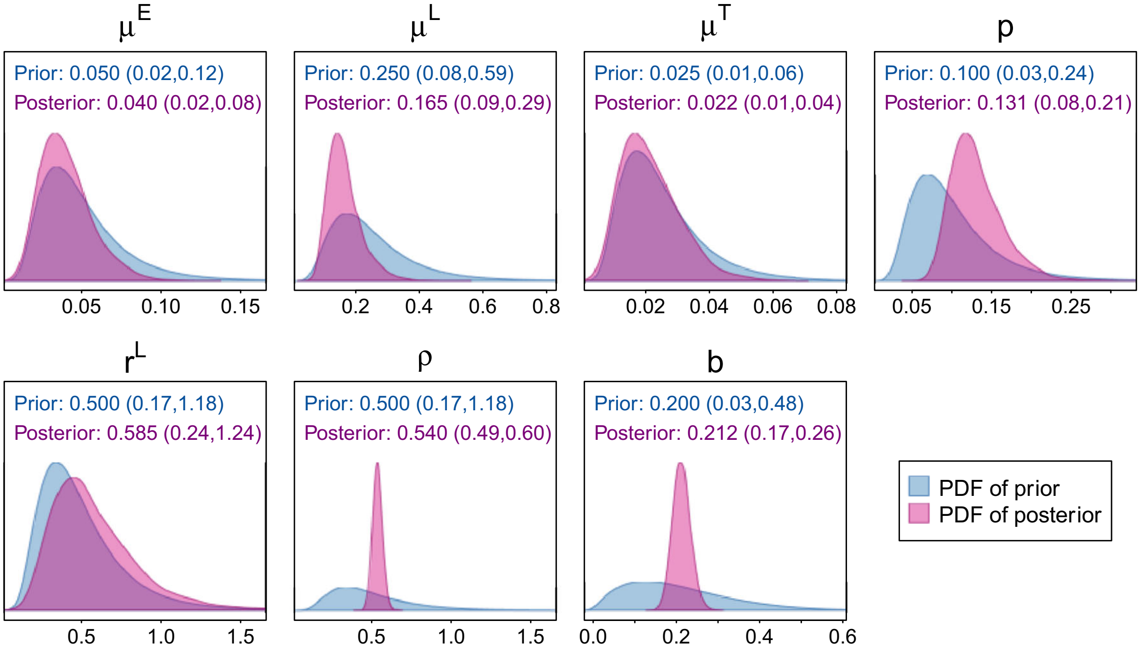


Fig. 3 Marginal prior and posterior distributions for model parameters included in the calibration. Text gives means of distribution, with equal-tailed 95% intervals shown in *parentheses*. *PDF* probability density function

Stochastic uncertainty may be less problematic when calculating the full posterior, as methods for sampling from the posterior distribution can be more robust to random error in the likelihood estimate. However, even in these situations it can be advantageous to reduce the magnitude of the Monte Carlo error. For example, to calibrate a microsimulation model of colorectal cancer, Rutter et al. estimated the parameters of the data likelihood by simulation, that is, estimating a large number of individual disease courses and using this sample to calculate a maximum likelihood estimator for the parameters of the calibration likelihood [26, 38]. The Monte Carlo error in these values will be inversely proportional to the square root of the size of this simulated sample, and thus a larger sample size reduces the magnitude of the error in the estimated likelihood. Another approach involves first estimating a smooth function (such as a spline or Gaussian process) to emulate the model and then using this function (which should be much quicker to evaluate) to calibrate the parameters [39]. Given differences in model complexity and computing resources, users of stochastic models may need to investigate a solution tailored to their situation.

3.6 Analytic Results Following Calibration

Having obtained a sample of calibrated parameter sets, the analysis proceeds as a conventional Monte Carlo simulation—the model is evaluated for the sample of calibrated parameter sets, and quantities of interest calculated from the distribution of modeled outcomes that are produced. As the treatment cost (c^T) was omitted from calibration, the sample of 10,000 parameter sets from IMIS was augmented by a sample of 10,000 values drawn from the prior for c^T. Using the parameter sets produced by IMIS, incremental life-years lived were estimated to be 130,000 [64, 228], and incremental costs were estimated to be US$123 million [−4, 312]. The ICER is calculated to be US$947 per life-year saved. This is lower than the ICER obtained from the uncalibrated model (US$1300) and uncertainty is substantially reduced, as shown in Fig. 5, which plots cost-effectiveness acceptability curves for uncalibrated and calibrated models.

4 Discussion

Calibrating simulation models to fit empirical data can improve the quality of model predictions and increase confidence among the consumers of study results [1, 40]. Bayesian methods provide a powerful approach for calibrating health policy models, with the model enabling a Bayesian evidence synthesis of many different sources of information relevant to the policy question [18]. These methods must be applied and interpreted thoughtfully, and the best approach for a particular application will depend on the model structure, purpose of the analysis, and availability/characteristics of calibration data. However, the general concepts described in this tutorial are

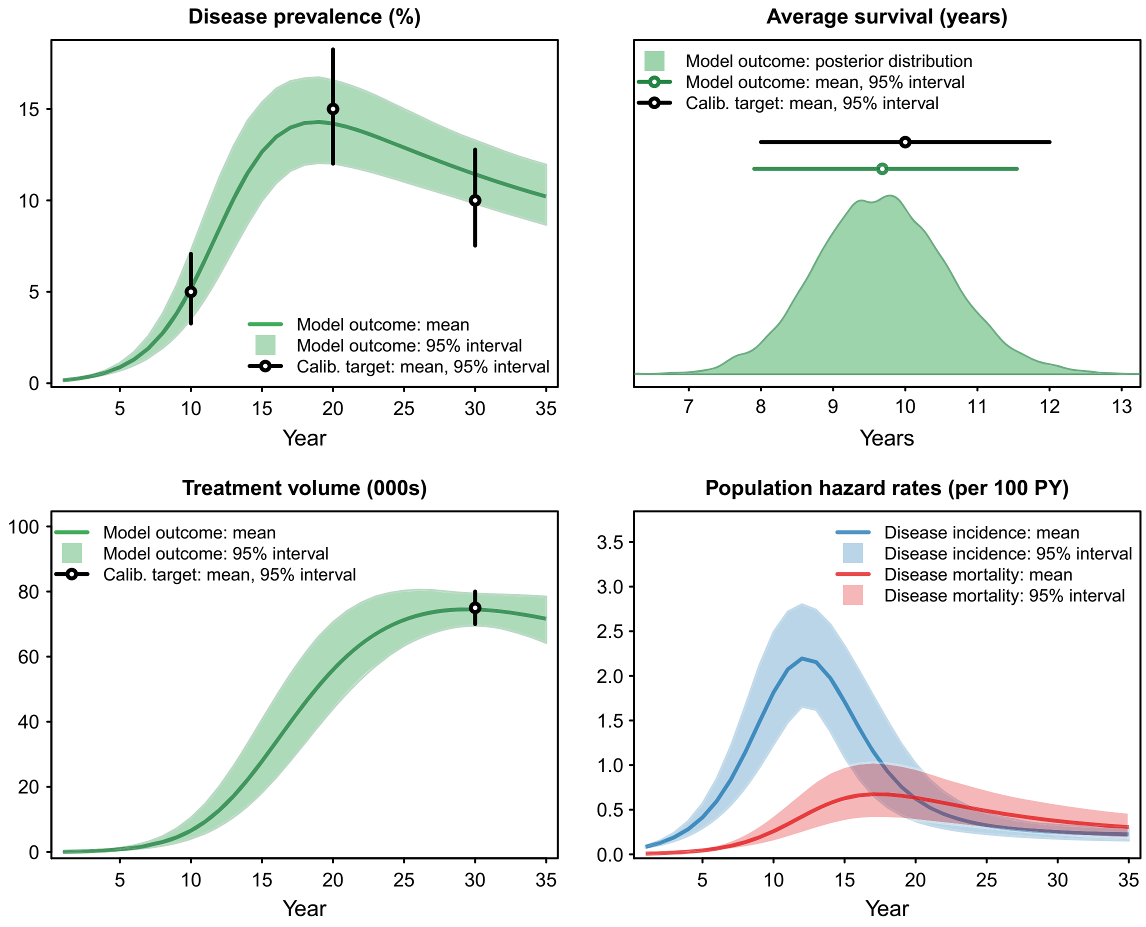
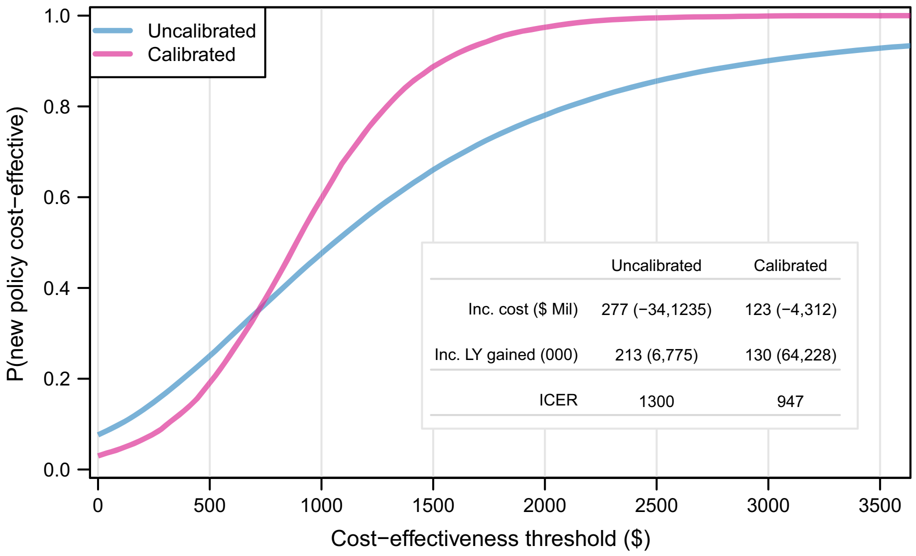


Fig. 4 Comparison of fitted model with calibration targets, and other modeled outcomes. For time-varying outcomes, means and intervals represent posterior means and equal-tailed 95% posterior intervals, respectively, calculated from the distribution of model results for each year of the simulation. *Calib.* calibration, *PY* person-years

Fig. 5 Cost-effectiveness acceptability curves for the uncalibrated and calibrated models. *ICER* incremental cost-effectiveness ratio, *Inc.* incremental, *LY* life-year



	Uncalibrated	Calibrated
Inc. cost ($ Mil)	277 (−34,1235)	123 (−4,312)
Inc. LY gained (000)	213 (6,775)	130 (64,228)
ICER	1300	947

applicable to a wide variety of models, and the underlying theory is agnostic to model structure provided the model is an adequate approximation of the system it was designed to represent. The worked example describes several techniques that can be employed for calibration, but does not cover the universe of algorithms and software that might be employed. Given the proliferation of Bayesian methods, this list is large and expanding, with contributions from fields such as engineering and machine learning [41, 42]. For regression-type models and relatively simple mechanistic models, available software can automate many aspects of calibration [27, 28]. Conversely, for models that are computationally expensive, or which involve substantial stochastic uncertainty, it may be difficult to identify an adequate approach. In these situations, individualized solutions may be required, and new methods tailored to these situations are needed [26, 39].

5 Conclusions

The theory underlying the Bayesian approach provides a principled framework for making analytic choices. However, in the course of specifying priors, likelihood, and model for a complicated policy problem, the analyst must make many decisions for which the underlying theory provides only general guidance. To the extent possible, the impact of these decisions should be investigated to assess their impact on study conclusions, and any disagreement between different sources of evidence examined. It may not be possible to resolve all disagreements in the evidence base, or reject all analytic options apart from the one chosen. Yet, invariably, the policy choice will not wait on all potential issues to be resolved. In this context, calibration should be considered an exercise in creating a reasonable model that produces valid evidence for policy—subject to ongoing scrutiny, revision, and improvement—rather than as a technique for identifying a unique theoretically optimal summary of the evidence [43].

Author Contributions N.A.M., J.J.K., A.P., and D.I.S. conceived the paper. N.A.M. wrote the paper. J.J.K., A.P., and D.I.S. edited the paper.

Compliance with Ethical Standards

Funding NAM received funding from National Institutes of Health/National Institute of Allergy and Infectious Diseases (Grant No. 7R01AI112438-02).

Conflict of interest The authors report no conflicts of interest directly relevant to the content of this article.

References

1. Vanni T, Karnon J, Madan J, et al. Calibrating models in economic evaluation: a seven-step approach. Pharmacoeconomics. 2011;29(1):35–49.
2. Enns EA, Cipriano LE, Simons CT, Kong CY. Identifying best-fitting inputs in health-economic model calibration: a Pareto frontier approach. Med Decis Making. 2015;35(2):170–82.
3. Waller LA, Smith D, Childs JE, Real LA. Monte Carlo assessments of goodness-of-fit for ecological simulation models. Ecol Model. 2003;164(1):49–63.
4. Wong RK, Storlie CB, Lee T. A frequentist approach to computer model calibration. J Royal Stat Soc Ser B (Stat Method). (In press).
5. De Finetti B. La Previsoin: ses lois logiques, ses sources subjectives. Annales de l'Institut Henri Poincare. 1937;7:1–68.
6. Ramsey FP. Truth and probability. In: Braithwaite RB, editor. Foundations of mathematics and other essays. London: Routledge & Keegan Paul; 1931.
7. Raiffa H, Schaifer R. Applied statistical decision theory. Boston: Harvard Business School; 1961.
8. van der Steen A, van Rosmalen J, Kroep S, et al. Calibrating parameters for microsimulation disease models: a review and comparison of different goodness-of-fit criteria. Med Decis Making. 2016;36(5):652–65.
9. Brown T, Grassly NC, Garnett G, Stanecki K. Improving projections at the country level: the UNAIDS Estimation and Projection Package 2005. Sex Transm Infect. 2006;82(Suppl. 3):iii34–40.
10. Gilbert JA, Meyers LA, Galvani AP, Townsend JP. Probabilistic uncertainty analysis of epidemiological modeling to guide public health intervention policy. Epidemics. 2014;6:37–45.
11. Weinstein MC, Siegel JE, Gold MR, et al. Recommendations of the panel on cost-effectiveness in health and medicine. JAMA. 1996;276(15):1253–8.
12. Briggs AH, Weinstein MC, Fenwick EA, et al. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. Med Decis Making. 2012;32(5):722–32.
13. Sanders GD, Neumann PJ, Basu A, et al. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. JAMA. 2016;316(10):1093–103.
14. Briggs AH. A Bayesian approach to stochastic cost-effectiveness analysis. Health Econ. 1999;8(3):257–61.
15. Fenwick E, O'Brien BJ, Briggs A. Cost-effectiveness acceptability curves: facts, fallacies and frequently asked questions. Health Econ. 2004;13(5):405–15.
16. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. Stat Methods Med Res. 2001;10(4):277–303.
17. Ades AE, Sculpher M, Sutton A, et al. Bayesian methods for evidence synthesis in cost-effectiveness analysis. Pharmacoeconomics. 2006;24(1):1–19.
18. Jackson CH, Jit M, Sharples LD, De Angelis D. Calibration of complex models through Bayesian evidence synthesis: a demonstration and tutorial. Med Decis Making. 2015;35(2):148–61.
19. Welton NJ, Ades AE. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. Med Decis Making. 2005;25(6):633–45.
20. Ades AE, Welton NJ, Caldwell D, et al. Multiparameter evidence synthesis in epidemiology and medical decision-making. J Health Serv Res Policy. 2008;13(Suppl. 3):12–22.
21. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2013.
22. Rubin D. Using the SIR algorithm to simulate posterior distributions. Bayesian Stat. 1988;3:395–402.
23. Kish L. Survey sampling. New York: Wiley; 1965.

24. Raftery AE, Bao L. Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. Biometrics. 2010;66(4):1162–73.
25. Whyte S, Walsh C, Chilcott J. Bayesian calibration of a natural history model with application to a population model for colorectal cancer. Med Decis Making. 2011;31(4):625–41.
26. Rutter CM, Miglioretti DL, Savarino JE. Bayesian calibration of microsimulation models. J Am Stat Assoc. 2009;104(488):1338–50.
27. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS: a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput. 2000;10:325–37.
28. Carpenter B, Gelman A, Hoffman M, et al. Stan: a probabilistic programming language. J Stat Softw. (In press).
29. Sunnaker M, Busetto AG, Numminen E, et al. Approximate Bayesian computation. PLoS Comput Biol. 2013;9(1):e1002803.
30. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162(4):2025–35.
31. Cowles MK, Carlin BP. Markov Chain Monte Carlo convergence diagnostics: a comparative review. J Am Stat Assoc. 1996;91(434):883–904.
32. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. J Comput Graph Stat. 1998;7(4):434–55.
33. Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica. 1996;6(4):733–60.
34. Gelman A. Exploratory data analysis for complex models. J Comput Graph Stat. 2004;13(4):755–79.
35. Garrett ES, Zeger SL. Latent class model diagnosis. Biometrics. 2000;56(4):1055–67.
36. Iman RL, Helton JC, Campbell JE. An approach to sensitivity analysis of computer models. Part II: ranking of input variables, response surface validation, distribution effect, and technique synopsis variable assessment. J Qual Technol. 1981;13:232–40.
37. Iman RL, Helton JC, Campbell JE. An approach to sensitivity analysis of computer models. Part I: introduction, input variable selection and preliminary variable assessment. J Qual Technol. 1981;13:174–83.
38. Rutter CM, Savarino JE. An evidence-based microsimulation model for colorectal cancer: validation and application. Cancer Epidemiol Biomark Prev. 2010;19(8):1992–2002.
39. Farah M, Birrell P, Conti S, De Angelis D. Bayesian emulation and calibration of a dynamic epidemic model for A/H1N1 influenza. J Am Stat Assoc. 2014;109(508):1398–411.
40. Stout NK, Knudsen AB, Kong CY, et al. Calibration methods used in cancer simulation models and suggested reporting guidelines. Pharmacoeconomics. 2009;27(7):533–45.
41. Kong CY, McMahon PM, Gazelle GS. Calibration of disease simulation model using an engineering approach. Value Health. 2009;12(4):521–9.
42. Cevik M, Ergun MA, Stout NK, et al. Using active learning for speeding up calibration in simulation models. Med Decis Making. 2016;36(5):581–93.
43. Gelman A, Shalizi CR. Philosophy and the practice of Bayesian statistics. Br J Math Stat Psychol. 2013;66:8–38.