Published in final edited form as:

Epidemics. 2021 December ; 37: 100514. doi:10.1016/j.epidem.2021.100514.

Bayesian validation framework for dynamic epidemic models

Sayan Dasgupta[*], **Mia R. Moore**, **Dobromir T. Dimitrov**, **James P. Hughes**

Fred Hutchinson Cancer Research Center, Seattle WA 98122, USA

Abstract

Complex models of infectious diseases are used to understand the transmission dynamics of the disease, project the course of an epidemic, predict the effect of interventions and/or provide information for power calculations of community level intervention studies. However, there have been relatively few opportunities to rigorously evaluate the predictions of such models till now. Indeed, while there is a large literature on calibration (fitting model parameters) and validation (comparing model outputs to data) of complex models based on empirical data, the lack of uniformity in accepted criteria for such procedures for models of infectious diseases has led to simple procedures being prevalent for such steps. However, recently, several community level randomized trials of combination HIV intervention have been planned and/or initiated, and in each case, significant epidemic modeling efforts were conducted during trial planning which were integral to the design of these trials. The existence of these models and the (anticipated) availability of results from the related trials, provide a unique opportunity to evaluate the models and their usefulness in trial design. In this project, we outline a framework for evaluating the predictions of complex epidemiological models and describe experiments that can be used to test their predictions.

Keywords

Epidemiological model validation; Markov Chain Monte Carlo; Bayesian credible interval; HIV transmission model

1. Introduction

Complex nonlinear simulation models have been widely used in applied scientific disciplines, including the study of climate, geophysics, soil and air pollution, epidemiology,

[*]Corresponding author. sdasgup2@fredhutch.org (S. Dasgupta).

CRediT authorship contribution statement

Sayan Dasgupta: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Mia R. Moore:** Writing – review & editing. **Dobromir T. Dimitrov:** Writing - review & editing. **James P. Hughes:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Software and code

The HIV transmission model and the software codes (in R version 3.5.1) for the analyses presented in Section 3 are available at the Github repository https://github.com/sdgroopkund/bayesian_validation_epimodels.git.

ecology, and other fields (see for example Caswell, 1976; Guttorp and Walden, 1987; Sampson and Guttorp, 1999; Poole and Raftery, 2000). Different types of models are available for this purpose, including stochastic, agent-based and deterministic differential equation models, the choice of which must be suited to the question of interest (see Halloran et al., 2017). While stochastic and agent based models incorporate random variation adding individual interactions such as partnership formation/dissolution, differential equation models are typically simpler and run faster, and are quite useful for analyses that require multiple runs of the model.

These models have also dominated the field of epidemic modeling, especially in complex diseases like HIV/AIDS (Johnson and White, 2011). In HIV epidemiology, these complex, dynamic models of infectious disease are used to understand the transmission dynamics of the disease, to obtain estimates and short-term projections of HIV epidemics (see Stanecki et al., 2012; Dorrington et al., 2005), to predict the effect of interventions (Case et al., 2018; de Montigny et al., 2018), to evaluate program implementation costs (Eaton et al., 2014; Schwartlander et al., 2011), to interpret clinical trial results (Adamson et al., 2019; Dimitrov et al., 2019) and to provide information for power calculations of community level intervention studies (Hayes and Moulton, 2017). For example, the increase in availability of antiretroviral therapy (ART) to individuals with HIV (WHO, 2013; Johnson et al., 2013) has opened up questions about the trajectory of the epidemic in presence of ART, including the probability of, and the level of coverage necessary for, complete eradication of the disease. To answer these questions, several models have been proposed to quantify the short and long term consequences of widespread ART availability and use on HIV prevalence, incidence and mortality, and to assess the level of coverage of ART necessary to substantially reduce or eliminate HIV (Granich et al., 2009; Kretzschmar et al., 2013).

Validity of results from a model is greatly dependent on the accuracy of the model structure and the input parameters, so a key goal during model development is to quantify and reduce the uncertainty about the structure and select model parameters which reproduce available epidemic data, using all of the available sources of evidence. This step is called calibration, and it involves careful tuning of the model structure to epidemiological endpoints, informed by available knowledge in the literature, and establishes credibility of the model (Ramin and Arhonditsis, 2013). Another important step in epidemiological model building is model validation, which is defined as a set of methods that help assess a model’s performance in making predictions. There are different types of validation processes, including face validity (the extent to which a model and its assumptions correspond to the current science as judged by experts), internal validity (whether the model behaves as intended and has been properly implemented), cross validity (how the model behaves when calibrated to one part of the data and used to predict the other), and external validity (to compare the predicted outcomes to measurable results). Each type of validation has its own sets of methods, strengths, limitations, and best practices (see Eddy et al., 2012).

For a clinical trial that was conducted based on insights gained from a model(s) in the planning phase, it can be quite useful to check how accurately the model(s) predicted the outcomes of interest when the trial is over. This is a possible external validation step. It is also quite useful to see if any source of bias in the model can be re-accounted for, and whether the model structure can be recalibrated to ensure its re-usability in the future (Eddy et al., 2012). The results of these analyses can help investigators determine applicability of the model in the public-health decision making process, and in planning of future trials.

In this article, we propose a framework for model comparison and evaluation based on observed data. To be precise, we are interested in external validation (will be called validation hereon) and re-calibration of a model, based on observed data. It is clear that the most powerful way to validate the prediction of a mathematical model is to record the model and its predictions in advance of new observations. Recently several community level randomized trials of combination HIV intervention have been planned and/or initiated (Boily et al., 2012). Significant epidemic modeling efforts were conducted during trial planning and became integral to the design of these trials. The existence of these models that have been designed to predict trial results in a specific setting, and the (anticipated) availability of results from those trials, provides a unique opportunity to evaluate those models and their usefulness in trial design through an external validation step. In this type of analysis, the goal is to test and potentially validate the model predictions of intervention with the trial results at the end of the trial. If the trial results are similar to the model predictions, then this validates the model for the trial in question, and the model can be used for further analyses with more confidence (Boily et al., 2012).

Interestingly, there have been relatively few such opportunities to rigorously evaluate the predictions of HIV epidemic models until now. Indeed, while there are multiple theoretical methods for calibration and external validation of complex models (see Bayarri et al., 2007; Kennedy and O’Hagan, 2001; Poole and Raftery, 2000), fairly simple procedures have been used for calibration and validation of models partially due to lack of substantial high quality, population-level data on HIV incidence, HIV prevalence and sexual behavior. The main goal of this article is to develop a step-by-step methodological framework for model validation, comparison and selection, when detailed data are available. Our approach takes into account the key uncertainties in model parameters, allowing for subsequent recalibration of the model, if needed and/or possible. An additional objective is to determine the extent to which external data can be used to reject a complex epidemiological model. The key components of the approach include the use of Gaussian process response-surface method, and introduction of Bayesian representations of model bias and uncertainty, following the works of Bayarri et al. (2007), Kennedy and O’Hagan (2001) and Kennedy et al. (2002).

In Section 2, we formulate the problem and propose our methodological framework for validation of a dynamic epidemic model based on the Bayesian validation framework of Bayarri et al. (2007). Additionally, we propose a way to make decisions regarding model fit based on Bayesian credible intervals. In Section 3, we describe a mathematical model (M_W) of a heterosexual HIV epidemic originally considered in Woods et al. (2018) to investigate how the proportion of early transmission affects the impact of ART on reducing HIV incidence. The model includes stages of HIV infection, flexible sexual mixing, and changes in risk behavior over the epidemic, and was calibrated to HIV prevalence data from South Africa. Next, we apply the proposed methodology to evaluate the performance of this model under departures of different model assumptions. In other words, we try to falsify this model under null and alternative conditions, and note how it performs in such situations. In

Section 4, we discuss the implications of our results. Some additional results are provided in Appendix.

2. Methods

2.1. Model framework

2.1.1. Epidemiological model definition—In this article, we consider epidemiological models of the form $y_M = M(x, \theta)$. Here $y_M \in \Omega_M$, the model output, typically emulates a population level outcome of interest, such as HIV prevalence. Each entry in y_M represent the projected outcome at different times within different demographics or different risk subgroups. The model M is a complex non-linear function, typically defined as a solution to a system of ordinary differential equations, which translates a vector of community-specific parameters $x \in \Omega_x$ and a vector of global parameters $\theta \in \Omega_\theta$ into the outcome y_M.

Throughout this paper a community will be any population within which an epidemic can be modeled independently from other communities. A community may be all individuals in a given city or region or may only be a subset, such as people who inject drugs, within which the epidemic may be evolving independently. A community may be divided into several interconnected sub-populations, for example divided by demography.

The community-specific parameters x (a vector of size p_x) denote information about attributes of a given community within which we wish to calculate the model output y_M. In the context of HIV prevention trials, y can be the projected HIV prevalence by gender and age group, and x might include community-specific information about level of ART usage, initial testing coverage (that differ by subgroups or communities), HIV prevalence for that site at some previous time point. The joint spaces Ω_x and Ω_M together represent the granularity of the model, the class of all possible sub-populations and communities for which the model M, after calibration, can predict the population-level outcome separately.

The ‘global’ parameters θ (of size p_θ) describe the part of the model that is universal, or fixed across different communities. For example, in the context of a HIV trial, θ contain various biologic, behavioral or intervention characteristics (e.g., parameters for intervention efficacy, transmission risk per act, sexual mixing, background level of male circumcision etc.) which are shared between community-specific simulations.

The components of y_M, x, and θ depend on the modeling context and intended granularity. For example, it may be crucial for a model on HIV transmission to predict the response separately for men who have sex with men (MSMs) and transgender women (TGW), two crucial sub-populations at risk for the disease. However that distinction may be inconsequential in the context of some other disease (say cancer), and a universal response may be preferred under a model aiming to predict the disease for these sub-populations. We define the ‘model structure’ of M to be the three spaces Ω_x, Ω_θ, and Ω_M.

2.1.2. Sources of error in the model—In principle, if x for a given sub-population and θ are known precisely, and the model is exactly ‘true’, then y_M or $M(x, \theta)$ should be able to predict the underlying truth without error. In practice, however, it is quite likely that there will be uncertainty regarding the true values of x and θ, and the model structure M is only a useful approximation of the underlying mechanics. Ideally, we would assess a model M by comparing the modeled epidemiological value y_M with the true value (denoted by y_R), though this may not be possible, as the reality will have to be estimated with error from a sample of the population. All these sources of uncertainties must be considered in assessing the model predictions.

2.1.3. Field data vs. reality—

Suppose we are interested in evaluating our model for K different subgroups, stratified under one or more baseline/longitudinal characteristics or communities, based on model projections at a given time during the study and the true outcome at that time. As mentioned before, we assume that the outcome is not observed directly, but it can be measured without bias. Let $y_{R,k}$ be the true outcome for community k, $k = 1,\dots,K$, at the specified time, and assume that we observe $y_{F,k}$, the 'field data' for this community at that time, as reality measured with error, that is,

$$y_{F,k} = y_{R,k} + \epsilon^F(y_{R,k}), \tag{1}$$

where $\epsilon^F(y_{R,k})$ are mean 0 errors with variance $1/\lambda_F(y_{R,k})$, for a precision process $\lambda_F(\cdot)$, which may or may not be explicitly known (we discuss this in details in Section 2.3.1). Further distributional assumptions can be made on $\epsilon^F(y_{R,k})$ under specific problem setups, for example, one operational assumption often used in validation problems is that $\epsilon^F(y_{R,k})$'s are independent normal random errors (see Bayarri et al., 2007). However, one might have to consider transformations of the responses $y_{R,k}$ and $y_{F,k}$ to enable easier distributional assumptions and interpretation of the error function $\epsilon^F(\cdot)$ in Eq. (1).

Note: Note that although y_F and y_M can be evaluated at several times during the study, we consider only one fixed time point here to compare the model projections with observed data, and as a result to simplify notations, we have dropped the suffix t from y_F, y_M as well as y_R.

2.1.4. Induced prior for y_M—

Information about different parameters of a model are often available from previously conducted trials or relevant observational studies in the literature. Knowledge from these various sources are combined for the quantification of the community-specific and global parameters, and the model is then carefully calibrated, resulting in well-informed prior distributions for these quantities. Thus, in this article, we assume that X_k,[1] the community-specific parameters for the kth subgroup, and Θ, the global parameter vector, are random variables with distributions $\mathscr{G}_{X_k}(\cdot\,;\eta_{X_k})$ and $\mathscr{H}_{\Theta}(\cdot\,;\eta_{\Theta})$ respectively, and with underlying hyperparameters η_{X_k} and η_Θ. These distributions quantify the level of uncertainties about the different model inputs, the community-specific and global parameters. Also note, that the distributions of X_k and Θ jointly induce a distribution, denoted by $\mathscr{F}^{M}_{X_k,\Theta}$, on the random model function $M(X_k, \Theta)$.

[1] We will use X_k, Θ etc to denote Random Variables, while x_k and θ will be reserved to denote observed realizations of such, and/or to denote fixed quantities.

Note: Simplification of the above setup is possible too, for example, if information about X_k is available through high quality estimates $\hat{x}_k$, we can assume that the true underlying community-specific parameter set is known *apriori*, and $X_k = \hat{x}_k$. Similarly, information about $\mathscr{H}_\Theta$ can be available only as scalars, $\Theta = \hat{\theta}$, that are available as estimates/calibrated values for the 'true' structural parameters of the model. In such cases, we may choose to use these scalars directly as 'true' (or best known) values in the analysis.

2.2. Model discrepancy

2.2.1. The discrepancy function, $\mathbf{D_M}$—Statistical assessment of a model prediction is based on the difference between the true value y_R and the model prediction. To that effect, we define the discrepancy function D_M, which we will use throughout this article. For a given model M with an instance of community-specific parameters x and global parameters θ, the discrepancy between the model output and the relevant response process y is given as,

$$D_M(y, x, \theta) = y - M(x, \theta)\,. \tag{2}$$

In our problem setup, as previously mentioned in Section 2.1.4, X_k, the community-specific parameters for the kth subgroup, and Θ, the global parameter vector, are random variables, while the true response for is $y_{R,k}$. Hence, the 'random' discrepancy function for subgroup k becomes $D_M(y_{R,k}, X_k, \Theta) = y_{R,k} - M(X_k, \Theta)$. We can analyze the distribution of $D_M(y_{R,k})$ for $k = 1, \ldots, K$ (for each of the different subgroups/communities) to evaluate model performance. The variation in $D_M(y_{R,k})$ between these communities can be useful in identifying areas where the model has failed to adequately capture the effect of factors on the response.

Note: Although we are interested in learning about $D_M(y_{R,k})$, in the absence of concrete knowledge about y_R, we typically have to rely on $D_M(y_{F,k})$ to infer about this quantity.

2.2.2. A Bayesian approach to evaluate $\mathbf{D_M}$—Since $D_M(X_k, \Theta)$ is a function of the random variables X_k and Θ, for which we have distributions of the form of $\mathscr{G}_{X_k}(\,\cdot\,; \eta_{X_k})$ and $\mathscr{H}_\Theta(\,\cdot\,; \eta_\Theta)$, one trivial way to obtain and analyze the distribution of D_M is through the transformation $D_M(X_k, \Theta) = y_{R,k} - M(X_k, \Theta)$ on the induced distribution $\mathscr{F}^M_{X_k, \Theta}$. Although the true response $y_{R,k}$ is typically unknown, one can replace $y_{R,k}$ by its observed estimate $y_{F,k}$. Although the resultant distribution is well-defined, it can only be used to assess the model in its current state. However, our interests may also lie in identifying the possible sources of error in the model, and if possible, recalibrate these faulty parts, based on observed data, to improve model performance.

Hence, using a Bayesian analysis described in the next section (Section 2.3), we propose to estimate the posterior distribution of the community-specific parameters X_k, global parameters Θ, and the discrepancy function $D_M(X_k, \Theta)$, to assess the model for community k and identify its sources of discrepancy. Ideally, the distribution of D_M should be centered around zero with small variance. We will show how to quantitatively evaluate this property, even in high dimensions using the idea of posterior tail probability (or p_{TP}). We will also show how to use this procedure to identify the different sources of error or bias in model parameters, both community-specific and global, and model structure.

2.2.3. Gaussian Stochastic Processes (GaSP) model as a prior for D_M—

We define a functional prior $P_{D_M}(x, \theta)$ on $D_M(x, \theta)$ for each pair ($X = x, \Theta = \theta$) using a Gaussian Stochastic Process (GaSP). A GaSP is useful in modeling functional outputs like D_M that depend smoothly on its arguments, x and θ (see Sacks et al., 1989; Currin et al., 1991). Although a GaSP is most helpful when all of the model arguments are continuous, it can also handle cases when one or more (but not all) of its arguments are discrete, by specifying a separate Gaussian response surface at each level of the discrete factors. This, combined with the priors on X and Θ, produces a prior for $D_M(X, \Theta)$.

Consider our setup of K subgroups (or communities), with community-specific random variables X_k and a global random variable Θ. Suppose we want to run the model M for $X_k = \tilde{x}_k$, for $k \in \{1,\ldots,K\}$, and $\Theta = \tilde{\theta}$, and evaluate the discrepancy of the model for these subgroups for the given realizations. Define the $K \times 1$ vector of discrepancies for the K subgroups as $\mathbf{d}_M = \left(D_M\left(\tilde{x}_1, \tilde{\theta}\right), \ldots, D_M\left(\tilde{x}_K, \tilde{\theta}\right)\right)^T$. Let Z_k denote the multivariate random variable $Z_k = (X_k, \Theta)$, such that $\mathbf{z}: = \left(z_1^T, \ldots, z_K^T\right)^T$ forms a $K \times p_z$ matrix, where $p_z = p_x + p_\Theta$ and $z_k = \left(\tilde{x}_k, \tilde{\theta}\right)$ is a realization of Z_K. The GaSP formulation assigns a K-variate Gaussian distribution for the vector $\mathbf{d}_M$ with mean function $\mu_d(\mathbf{z}) = (\mu_d(z_1), \ldots, \mu_d(z_K))^T$. The covariance function $C_d(\cdot,\cdot)$ for the Gaussian process is given as,

$$C_d(z_k, z_l) = \frac{1}{\lambda_d}\exp\left(-\sum_{j=1}^{p_z} \beta_{d,j}\left|z_{k,j} - z_{l,j}\right|^{\alpha_{d,j}}\right) \tag{3}$$

for parameters $\beta_d = \left\{\beta_{d,1}, \ldots, \beta_{d,p_z}\right\}$ and $\alpha_d = \left\{\alpha_{d,1}, \ldots, \alpha_{d,p_z}\right\}$ that control the amount of correlation between $D_M(z_k)$ and $D_M(z_l)$, and with λ_d controlling the precision of the Gaussian surface. Note that for elements $z_k = \left(\tilde{x}_k, \tilde{\theta}\right)$ and $z_l = \left(\tilde{x}_l, \tilde{\theta}\right)$ (where $k, l \in \{1,\ldots,K\}$), the last p_Θ elements in the sum $\sum_{j=1}^{p_z} \beta_{d,j}\left|z_{k,j} - z_{l,j}\right|^{\alpha_{d,j}}$ are 0, as $z_{k,j+p_x} = z_{l,j+p_x} = \tilde{\theta}_j$ for $j = 1,\ldots,p_\Theta$. Hence the above simplifies to $C_d(z_k, z_l) = \frac{1}{\lambda_d}\exp\left(-\sum_{j=1}^{p_x} \beta_{d,j}\left|x_{k,j} - x_{l,j}\right|^{\alpha_{d,j}}\right)$ and as a result only the first p_x elements of the hyperparameters β_d and α_d need to be specified.

Now let us briefly discuss the specifications of the hyperparameters μ_d, λ_d, α_d and β_d. First, to limit the number of hyperparameters in the model, the components of α_d are fixed. For example, $\alpha_d = 2$ gives us the standard Gaussian process formulation (see Bayarri et al., 2007), and for the rest of the article, we will consider the elements of α_d to be fixed at that value. The mean function of the GaSP, $\mu_d(\cdot)$, can be chosen to be fixed at a constant μ, the most intuitive choice being $\mu = 0$. To generalize the analysis further, we can specify symmetric mean 0 distributions to be used as priors if required. For elements of β_d, we can either specify priors for them though another set of hyperparameters, or fix them at estimates, determined using data driven methods. However, the precision parameter λ_d is typically kept stochastic, and we can specify priors like inverse gamma for λ_d with the hyperparameters (shape and scale) determined in a data-driven manner. Please see a more detailed discussion on how to choose these hyperparameters for a practical setting as our simulation example in Section 3.2.

2.3. Assessing the model

2.3.1. Specification for λ_F in the Bayesian analysis—

Often $\lambda_F(y_{R,k})$ can be determined by the sampling design of the field data $y_{F,k}$ for a given subgroup or community k. Consider the case of a epidemiological model that predicts HIV prevalence, where reality $y_{R,k}$ is a proportion in the interval (0,1), and is estimated through sampling. Now if the observed 'field data' y_F is collected based on a set of N samples, then it will be distributed as $Ny_{F,k} \sim \mathrm{Bin}(N, y_{R,k})$. Here the precision parameter λ_F has an explicit expression,

$$\lambda_F(y_{R,k}) = \frac{1}{Var(y_{F,k})} = \frac{N}{y_{R,k}(1-y_{R,k})}. \tag{4}$$

Note that for a large enough sample size, $y_{F,k}$ is approximately distributed as Gaussian with variance $1/\lambda_F(y_{R,k})$. A natural estimate for $\lambda_F(y_{R,k})$ can be produced by replacing $y_{R,k}$ in (4) by its estimate $y_{F,k}$. However in case the error structure is unknown or the estimates are unreliable, standard priors like $1/\lambda_F$ can be used if replicates are available, otherwise data dependent priors centered at a suitably defined estimate of $\lambda_F(y_{R,k})$ can be used in the analysis instead.

2.3.2. Calculation of posterior of D_M—

Using the multivariate (K-variate) prior for D_M, the induced model prior $\mathscr{F}^M_{X_k,\Theta}$, and the error process λ_F, we evaluate the posterior distribution of D_M at each of the K communities. Below we give details for the full Bayesian analysis. This analysis can be further modified based on different distributional assumptions on the various parts of the model (community-specific and/or global) and the observed data (error variance). One such modified analysis has been discussed in the Appendix, which was used in our simulation exercise. From here on, we will also assume that $\epsilon^F(y_{R,k})$'s are independent normal random errors. This assumption is certainly valid in many cases of epidemiological modeling, when model predictions are of standard forms (for example, binomial proportions like prevalence or rates like log-incidence) and the field data estimates are obtained using survey data or other standard methods, as the estimation error is approximately Gaussian due to the Central Limit Theorem. This is the case in our simulation example that we will study in detail later.

Suppose we want to run the analysis for K different communities with random community-specific parameters for the kth community, X_k, with $X_k \sim \mathscr{G}_{X_k}$, and a random global parameter set Θ, with distribution $\mathscr{H}_\Theta$. Recall the equations that connect $y_{R,k}$, $y_{F,k}$ and M; for community k,

$$y_{F,k} = y_{R,k} + \epsilon^F(y_{R,k}),$$

$$y_{R,k} = M(X_k, \Theta) + D_M(X_k, \Theta),$$

$$\text{where } \epsilon^F(y_{R,k}) \sim N(0, 1/\lambda_F(y_{R,k}))\,. \tag{5}$$

Also, suppose that the precision process is unknown, and we have prior distributions of the form $\lambda_F(y_{R,k}) \sim P(\lambda_F | y_{R,k})$ for community k. Given the unknowns, these produce a multivariate normal density for the collection of all field data $y_{F,k}$ for $k \in \{1,\ldots,K\}$, denoted by $f(y_{F,1}, \ldots, y_{F,K} \mid \Theta, X_1, \ldots, X_K, \lambda_F(y_{R,1}), \ldots, \lambda_F(y_{R,K}), D_M(X_1, \Theta), \ldots, D_M(X_K, \Theta), \lambda_d, \mu_d, \beta_d, \alpha_d)$. Removing the known (or fixed) quantities from the likelihood, the posterior density of the unknowns given the data $y_{F,k}$ for $k \in \{1,\ldots,K\}$ can be written as

$$\begin{gathered} P(\Theta, X_1, \ldots, X_K, \lambda_F(y_{R,1}), \ldots, \lambda_F(y_{R,K}), D_M(X_1, \Theta), \ldots, D_M \\ \times (X_K, \Theta), \lambda_d \mid y_{F,1}, \ldots, y_{F,K}) \\ \propto f(y_{F,1}, \ldots, y_{F,K} \mid \Theta, X_1, \ldots, X_K, \lambda_F(y_{R,1}), \ldots, \lambda_F(y_{R,K}), \\ D_M(X_1, \Theta), \ldots, D_M(X_K, \Theta), \lambda_d) \\ P(\Theta, X_1, \ldots, X_K, \lambda_F(y_{R,1}), \ldots, \lambda_F(y_{R,K}), \\ D_M(X_1, \Theta), \ldots, D_M(X_K, \Theta), \lambda_d)\,. \end{gathered} \tag{6}$$

Note that the prior $P(\Theta, X_1, \ldots, X_K, \lambda_F(y_{R,1}), \ldots, \lambda_F(y_{R,K}), D_M(X_1, \Theta), \ldots, D_M(X_K, \Theta), \lambda_d)$ can be written as

$$\begin{aligned} & P(\Theta, X_1, \ldots, X_K, \lambda_F(y_{R,1}), \ldots, \lambda_F(y_{R,K}), D_M(X_1, \Theta), \ldots, D_M(X_K, \Theta), \lambda_d) \\ & = \mathscr{H}_\Theta \mathscr{G}_{X_1} \ldots \mathscr{G}_{X_K} P(\lambda_F \mid y_{R,1}) \ldots P(\lambda_F \mid y_{R,K}) P(\lambda_d) P(\mathbf{D}_M \mid \Theta, X_1, \ldots, X_K, \lambda_d), \end{aligned} \tag{7}$$

where $P(\mathbf{D}_M \mid \Theta, X_1, \ldots, X_K, \lambda_d)$ is the multivariate GASP prior for the discrepancy at all K communities. From here onward, we will refer to the k-variate discrepancy random variable as $\mathrm{D}_M := \{D_M(X_1, \Theta), \ldots, D_M(X_K, \Theta)\}$.

The posterior distribution can be determined through MCMC techniques (Robert and Casella, 2004). The MCMC step can be performed in a number of different ways. In our simulation study, we have used the Metropolis Hastings (MH) algorithm to calculate the posterior distribution, however alternative techniques like Bayesian Melding (see Poole and Raftery, 2000), Hamiltonian Monte Carlo or HMC (see Chatzilena et al., 2019; Betancourt, 2017) can be used instead. For example, in comparison with the traditional MH algorithm, HMC can offer greater computational efficiency, especially in higher dimensional or more complex modeling situations. Please also see Robert and Changye (2020) for a comparative analysis of these methods to determine the most optimal framework for a given problem. In the Appendix, we describe the MH procedure for a simplified setup, which will be the setup for the simulation example presented in Section 3.

2.3.3. Identifying sources of error—The Bayesian procedure employed to obtain the posterior distribution of discrepancy vector $\mathbf{D}_M = \{D_M(X_1, \Theta), \ldots, D_M(X_K, \Theta)\}$ can be run in a number of different ways depending on the targeted goals of the validation procedure, as described below. We first describe what we mean by ‘updating a prior’ in this context. Note that the main goal of our Bayesian analysis is to obtain the posterior distribution of $\mathbf{D}_M$. This posterior distribution is informed by the priors on $\mathbf{D}_M$, the model structure M, the stochastic distributions $\mathscr{G}_{X_k}$ on X_k for $k \in \{1,\ldots,K\}$ and $\mathscr{H}_\Theta$ on Θ, and the observed data $y_{F,k}$ for $k \in \{1,\ldots,K\}$. Now, in the process of obtaining the posterior distribution of $\mathbf{D}_M$, we can additionally choose to obtain the posterior distributions of any, some or all of the random variables $X_1,\ldots,X_K,\Theta$. For parameters that we choose to update, their stochastic distributions are considered as priors and the relevant posteriors are obtained through the MCMC procedure. Parameters which are not being updated are not considered directly in the MCMC procedure. While, we do have to generate realizations of these parameters from their current stochastic distributions for calculating the likelihood in each MCMC update. Their distributions are not updated in the MCMC steps. In the Appendix, we have described how to achieve this in our simulation setup through the Metropolis Hastings procedure. Although the flow chart in Fig. 1 shows the flow of the analysis when the objective is to re-calibrate all parameters (if necessary to obtain a better model fit), Researchers may be interested in running only part of the analysis flow depending on study objectives. Below we present a few such ways to run the analysis:

1. **Evaluation of current model** To assess the model in its current state, we can run the analysis without updating the priors $\mathscr{H}_\Theta$ or $\mathscr{G}_X$. As a result we only obtain the posterior distribution of $\mathbf{D}_M$ along with the posteriors of the hyperparameters controlling the discrepancy distribution in the model.

2. **Recalibration of community-specific parameters** If we are interested in assessing the model after allowing for recalibration/re-tuning of the community-specific parameters, we can update $\mathscr{G}_{X_k}$ to obtain $\mathscr{G}_{X_k}^{\text{post}}$, the posterior distribution of X_k along with the posterior distribution of $\mathbf{D}_M$. Note that one might choose to update all of the community-specific parameters at a time together, or one by one, or in groups of more than one.

3. **Recalibration of community-specific and global parameters** If we are interested in assessing the model after allowing for recalibration of both the global and the community-specific parameters, we can choose to update $\mathscr{H}_\Theta$ to obtain $\mathscr{H}_\Theta^{\text{post}}$, the posterior distribution of Θ as well, along with $\mathscr{G}_{X_k}^{\text{post}}$, the posterior distribution of X_k and the posterior distribution of $\mathbf{D}_M$. Like in step 2, one might choose to update any combination of community-specific and global parameters at a time together, or in steps (for which new data become available after the last model evaluation), or till they are all updated.

Note: The entire analysis can be conducted in a single step or in a series of several steps, with each step devoted to updating and analyzing a specific parameter or a group of parameters. The order in which the community-specific and global parameters are updated in the analysis above is interchangeable, depending on the context and necessity. For example, the order can be determined by running sensitivity analyses to gauge the influence of each parameter (or parameter set) on the final model outputs, and then updating them in order of importance.

After running the analysis in one of the aforementioned ways, the posterior distribution of the discrepancy, $\boldsymbol{P}^{\text{post}}(\mathbf{D}_M)$, is assessed carefully to see if it is considerably removed from being centered at 0, which may suggest (depending on the analysis run) that any or all of (i) the structure of the model, (ii) the priors for Θ and (iii) the priors for X_K need to be updated. A step by step flowchart for the entire scope of the analysis under the discussed setup is given in Fig. 1.

2.3.4. Quantification of model performance with Posterior Tail Probability ($\mathbf{p_{T\,P}}$)

—When the model assumptions are satisfied, that is, when we have informative and precise priors on $X_1,\ldots,X_K,\Theta$ and the model is a close approximation of the reality, we expect the posterior distribution of the discrepancy to be roughly centered around $\mathbf{0}$. However, under violation(s) of any one or more of these assumptions, the posterior distribution will either suffer a shift in mean, or show inflated variance, be multimodal, or any combination of the above. In this section, we will concern ourselves only with the first type of alternative, that is, when the posterior distribution suffers a mean shift. The goal is to quantify the extent of this shift as a model validation metric. Note that our interests may lie in validating the model for (i) all K communities together, (ii) for a given subset of these K communities, or (iii) for each community separately. We present methods for the second situation here, and it is easy to see that the first and the third situations can be viewed as special cases of the second.

Let us assume that we want to assess the discrepancy of the model M at J out of K communities, $\left\{X_{i_1}, X_{i_2}, \ldots, X_{i_J}\right\}$ with $i_1, i_2, \ldots, i_J \in \{1,2,\ldots,K\}$ and $1 \le J \le K$. Let $\mathbf{D}_M$ be the multivariate random variable (of size $J \times 1$) representing discrepancy at these communities. The Mahalonobis distance for a discrepancy vector $\mathbf{D}_M = \mathbf{d}$ from a given center μ and a given covariance matrix V is

$$\Delta_{\mu,V}(\mathbf{d}) = \sqrt{(\mathbf{d}-\mu)'V^{-1}(\mathbf{d}-\mu)} \qquad (8)$$

The Posterior Tail Probability ($p_{T\,P}$) measures the probability that an observation drawn at random from the posterior distribution of $\mathbf{D}_M$ lies further away from the mean of that distribution than does $\mathbf{0}$, the expected discrepancy under the null hypothesis that the Model is true. Effectively, it is a (two-sided) tail probability, calculated based on the relative location of $\mathbf{0}$ with respect to the posterior distribution of $\mathbf{D}_M$.

Let $\boldsymbol{\eta}$ and V be the posterior mean and covariance matrix for the posterior distribution for $\mathbf{D}_M$, and $\bar{d}$ is the sample mean of discrepancies from the MCMC analysis, while $\widehat{V}$ is a sample estimate for V. Now note that the Mahalonobis distance for a discrepancy sample $\mathbf{D}_M = \mathbf{d}$ from the center $\boldsymbol{\eta}$ and for the given covariance matrix V is given as

$$\Delta_{\eta,V}(\mathbf{d}) = \sqrt{(\mathbf{d}-\eta)'V^{-1}(\mathbf{d}-\eta)}\,.$$

To estimate p_{TP} from the MCMC data, $\mathbf{d}_j$, we:

1. Calculate $\boldsymbol{\delta} = \{\delta_1,\ldots,\delta_N\}$, where δ_j is the distance between sample $\mathbf{d}_j$ and $\bar{d}$. That is, $\delta_j = \sqrt{\left(\mathbf{d}_j - \bar{\mathbf{d}}\right)'\hat{V}^{-1}\left(\mathbf{d}_j - \bar{\mathbf{d}}\right)}$.
2. Compute δ^0, the distance between $\mathbf{0}$ and $\bar{\mathbf{d}}$, as $\delta^0 = \sqrt{\bar{\mathbf{d}}'\hat{V}^{-1}\bar{\mathbf{d}}}$.
3. Calculate $\hat{p}_{TP} = \sum_{j=1}^{N}\left(\delta_j \ge \delta^0\right)/N$.

Note that p_{TP} estimates the probability that $\Delta_{\eta,V}(\mathbf{D}_M) \ge \Delta_{\eta,V}(\mathbf{0})$, that is, $\hat{p}_{TP} = \mathbb{P}(\Delta_{\bar{\mathbf{d}},\hat{V}}(\mathbf{D}_M) \ge \Delta_{\bar{\mathbf{d}},\hat{V}}(\mathbf{0}))$. Fig. 2 gives us an overview of how the posterior tail probability might look like when discrepancy is univariate (for a single community). Under H_0, the true center (mean) of the posterior discrepancy is 0, hence the distance between the true center and 0 is also trivially 0, that is, $\Delta_{\eta=0,V}(0) = 0$. Thus for any sample, its Mahalonobis distance from the true center (or its mean) will be greater than or equal to the distance between the true center and 0, and its posterior tail probability will be 1 (the probability under the shaded region A in the leftmost figure in Fig. 2). On the other hand, under a hypothetical alternative situation when the model is not valid (H_1) and that the true center (or mean) is at K, the Mahalonobis distance between a randomly chosen sample and the center will be greater than that between the center and 0 if the sample belongs to either of the shaded regions A and C, and will be lesser than that between the center and 0 if it belongs to region B.

Now, observe from Fig. 2, that the area of the region B is exactly $1 - p_{TP}$, and that the interval that bounds it forms a credible interval of level p_{TP} for the posterior discrepancy around its posterior mean, that is, the interval (0,2K) is a $C_{p_{TP}}$ for the posterior discrepancy $\mathbf{D}_M$. Mahalonobis distance helps to extend this idea to multivariate discrepancy vectors, where p_{TP} gives us the minimum α level of a credible region centered around the multivariate posterior mean that does not contain $\mathbf{0}$.

Because the MCMC chain consists of autocorrelated samples, we estimate $\hat{V}$ using an approach based on batch means Wakefield (see Chapter 3 of 2013), Glynn and Iglehart (see Chapter 3 of 1990). We split the output of N MCMC samples into L batches each of length I, with I chosen large enough that the batch means have low serial correlation, and then estimate V using the variance of the batch means. In our simulation examples, the number of MCMC samples, N, is 5000 and thus we have used $L = 50$ and $I = 100$. The procedure is given below in steps:

1. Compute the mean of the function of interest (for us, it is the vector of discrepancies at J communities) within each batch. That is, for $l = 1,\ldots,L$,

$$\hat{\mu}_l = \frac{1}{I}\sum_{i=(l-1)I+1}^{lI} \mathbf{d}_i$$

2. The overall mean is then given as $\hat{\mu} = \frac{1}{L}\sum_{l=1}^{L}\hat{\mu}_l$.

3. Note that $\sqrt{I}(\hat{\mu}_l - \hat{\mu}), l = 1, \ldots, L$ are approximately independently distributed as $N_J(\mathbf{0}, V(\mathbf{D}_M))$.

4. Thus, $V(\mathbf{D}_M)$ can be estimated by V as

$$\hat{V} = \frac{I}{L-1} \sum_{l=1}^{L} (\hat{\mu}_l - \hat{\mu})^2 .$$

3. A simulation study

In this section, we study our formulation in the context of HIV epidemic models. We start off this section by describing the model, and then we discuss the simulation settings and the results.

3.1. An HIV transmission model

Antiretroviral therapy (ART) has been shown to reduce the infectiousness of HIV infected persons, but only after HIV testing and diagnosis, linkage to care, and successful viral suppression (Hallett et al., 2009; Johnson et al., 2013). Thus a large proportion of HIV transmissions may occur during a period of high infectiousness in the first few months after infection, and is perceived as a threat to the impact of HIV "treatment-as-prevention" strategies. Woods et al. (2018) considered a mathematical model (M_W) of HIV epidemics in men through heterosexual contacts, to investigate and explore the population health implications of investing in evidence generation activities such as clinical trials, surveillance programs and health system performance measurement. The model is based on features of a generalized HIV epidemic in sub-Saharan Africa (SSA), and thus can be used as a simplified representation of HIV epidemiology in SSA to evaluate the HIV prevention and treatment strategies available there. More specifically, the model projects how the proportion of early transmission affects the impact of ART on reducing HIV incidence. It includes different stages of HIV infection, sexual mixing, and changes in risk behavior over the epidemic.

The model M_W (see Fig. 3) simulates HIV prevalence in a hypothetical population since 1980, in six different communities (see model diagram in Fig. 3). Prevalence of HIV in each community in 2013 forms the (scalar) community-specific parameter x_k for that community ($k \in \{1,\ldots,6\}$), given as $\mathbf{x}_0 = \{0.11,0.35,0.001,0.03,0.24,0.15\}$. The model estimates HIV transmission in each community, with parameters fit so that the prevalence in the model in 2013 is as specified in the input prevalence data (the community-specific parameter). If any intervention is applied within the study period, the model can project the effect of this intervention over the next period, compared to the scenario when no intervention is applied. In our hypothetical case, we assume that a combination intervention program is initiated in 2015, and the HIV prevalence is simulated up to 2030. The interventions that are modeled through this combination package in the model are (i) enhanced antiretroviral therapy (ART), which reduces the likelihood that a positive individual will transmit infection and the mortality rate from late-stage infections (ii) behavioral interventions (for example, counseling on condom use or other safe sex practices), which acts to modulate the force of infection; and (iii) medical male circumcision (MMC), which reduces the risk of men to acquire HIV infection. Model equations can be found in the Appendix (and also in Woods et al., 2018). There are 10 global parameters (same for all communities) in the model integrated in the model structure, namely,

1. The reduction in the rate of transmission from individuals on ART, ε_A.
2. The reduction in the risk of acquisition of HIV in circumcised men, ε_C.
3. The rate of leaving the population (due to aging out or death due to caused unrelated to HIV), μ_1.
4. The population growth rate, ε.
5. The rate of progression from infected state $I1$ to $I2$, σ_1.
6. The mortality rate due to HIV in late stage positive individuals in $I2$, μ_2.
7. The mortality rate due to HIV for individuals on ART, μ_3.
8. The rate of transmission of HIV (per partnership), $\beta(t)$, that controls the forces of infection, λ_1 and λ_2.
9. Background prevalence rate of HIV in the population in 1980, p_{ini}, considered same for all communities.
10. The proportion of people assumed to be already circumcised in 2013, same for each community, c.

The parameter τ, denoting the proportion of men entering the population at a given community who are not at risk, is fitted (using least squares) separately for each community, based on the population dynamics of that community at the time when the community-specific parameter was recorded (in our case prevalence data for that community in 2013). Note that τ is not considered as community-specific or global parameter, as it is not an input for the model but rather it is fitted based on the input (community-specific and global) parameters. The parameters ε_A and ε_C directly affect the not-at-risk parameter τ, as well as the forces of infection, λ_1 and λ_2. The prevalence of HIV, as well as progression and mortality rates due to the disease at a given time t is determined by the (time varying) rate of transmission parameter, $\beta(t)$, and the proportion of the population who are not at risk, τ. As mentioned before, the model was originally calibrated to HIV epidemics data from SSA, and the best point estimates for these parameters were calculated. We denote this estimate vector by θ_0.

3.2. The analysis setup

Before we can apply the Bayesian validation procedure on M_W, it is important to correctly define the response of interest, one which can give us a substantial idea about the fit of the model. In this analysis, the response was chosen to be the HIV prevalence in each community at the beginning of the year 2020, 5 years after the interventions were initiated in these communities. Next we need to define the distributions $\mathscr{G}_{X_k}$ and $\mathscr{H}_{\Theta}$ for X_k, $k = 1,\ldots,6$, and Θ respectively. To simplify things, we assume that there is no estimation error in $\mathbf{x}_0 = \{x_1,\ldots,x_6\}$, that is, the available information on the HIV prevalence in 2013 in these 6 communities reflect the actual truth, or in other words, $X_k := x_k$. We next define priors on the model parameters. Since all of the parameters are proportions between 0 and 1, we consider a logit-Gauss prior in our analysis, that is, logit(Θ) ~ $N(\mu, \sigma^2)$, with μ := logit(θ_0), and $\sigma^2 := p(\theta_0(1 - \theta_0))$, for a given scale parameter p.

For a given instance $\Theta := \theta$ and a given timepoint within 1980 and 2030, the model M_W outputs the HIV prevalence vector $\{M_W(x_1, \theta), \ldots, M_W(x_6, \theta)\}$. We assume that HIV prevalence in 2020 for community k can be estimated through a simple random sample of size n_k drawn from the population in question. Thus, given a true HIV prevalence of y_k at community k in 2020, we have an unbiased estimate $\hat{y}_k$, which can be written as $\hat{y}_k = \frac{Z_k}{n_k}$, where $Z_k \sim B(n_k, y_k)$ with sampling variability $Var(\hat{y}_k) = \frac{y_k(1 - y_k)}{n_k}$. We chose a sample size n_k = 2000 for each community in our simulations. Since prevalence is a proportion between 0 and 1, the HIV prevalence values are logit transformed, so that $D_M(x, \Theta)$ is distributed over the entire real line. We assume Eq. (5) from Section 2 hold with M = logit(M_W) and $X_k := x_k$, where,

1. $M(x_k, \theta) = \text{logit}(M_W(x_k, \theta))$,
2. $y_{R,k} = \text{logit}(y_k)$,
3. $y_{F,k} = \text{logit}(\hat{y}_k)$,

for k = 1,…,6. Note that due to this transformation, $\lambda_F(y_{R,k}) = 1/Var(y_{F,k}) = n_k \frac{e^{y_{R,k}}}{\left(1 + e^{y_{R,k}}\right)^2}$, and also that $\widehat{\lambda_F(y_{R,k})} = \lambda_F(y_{F,k})$.

As discussed in Section 2.2.3, we assume a GaSP prior for the discrepancies $\mathbf{D}_M$. For the Bayesian analysis, we can specify priors on the GaSP hyperparameters, μ_d, β_d and λ_d, as well. However, due to limited data availability, and owing to the fact that no direct data about model discrepancies are available, introducing too many hyperparameters in the model can increase collinearity between the GaSP parameters. Thus, we allow only λ_d to be stochastic, and fix the rest of the hyperparameters at reasonable values, following recommendations of Bayarri et al. (2007), and as discussed in Section 2.2.3. This is achieved in the following manner: the vector $\{y_{F,1} - M(x_1, \theta_0), \ldots, y_{F,6} - M(x_6, \theta_0)\}$ is treated as a realization from a multivariate normal with constant mean vector μ_d and covariance matrix $C_d(z_i, z_j)/\lambda_d + \Lambda_F^{-1}$, where $z_i = (x_i, \theta_0)$ and Λ_F is a diagonal matrix with kth diagonal entry $\lambda_F(y_{F,k})$. We use standard GaSP fitting software to obtain MLE estimates $\hat{\beta}_d$, that are used as the fixed value for β_d in this analysis. The mean function μ_d is fixed at 0. We assume an inverse gamma prior for λ_d with shape parameter $\alpha_{\lambda_d} = 1$ and scale parameter $\beta_{\lambda_d} = 5\hat{\lambda}_d$ where $\hat{\lambda}_d$ is the MLE estimate of λ_d obtained from the GaSP analysis described above, following suggestions from Bayarri et al. (2007). For the Metropolis Hastings algorithm, we also specify proposal distributions for unknown random variables, namely,

- A Gaussian distribution for $D_M(x, \Theta)$.
- A Beta distribution for Θ (since all are quantities between 0 and 1).

- An Inverse Gamma distribution for λ_d.

3.3. The simulation scenarios

We evaluate the model M_W under different scenarios, given by the following:

1. **Setting 1: The Null scenario:** The model is completely accurate, that is, both the model structure M and the parameter set θ_0 are accurate, that is, $y_{R,k} = M(x_k, \theta_0)$, for $k = 1,\ldots,6$. The priors $\mathscr{H}_\Theta(\cdot\,;\eta_\Theta)$ with hyperparameters $\eta_\Theta = \{\theta_0, \eta_\Theta^*\}$ reflect a valid quantification of the uncertainty regarding Θ. Plot for this setting is given in Fig. 4.

2. **Setting 2: Faulty prior information on Θ:** The model structure M is accurate, but information collected on one or more of the parameters is erroneous. To create this scenario, we start off with the assumption that θ_0 is still the true parameter vector, but we calibrated the model at an incorrect value θ_0^*, which is created by perturbing some of the elements of θ_0. The priors are now given by $\mathscr{H}_\Theta(\cdot\,;\theta_0^*, \eta_\Theta^*)$. In this scenario the following holds for each community k.

$$\begin{aligned} y_{R,k} &= M(x_k, \theta_0) \\ &= M\left(x_k, \theta_0^*\right) + M(x_k, \theta_0) - M\left(x_k, \theta_0^*\right) \\ &= M\left(x_k, \theta_0^*\right) + D^1\left(x_k, \theta_0^*\right) \end{aligned}$$

   In this case, we should be able to falsify the model with running only step 1 of Section 2.3.3. Running step 3 however should recalibrate the parameter set. We can evaluate the accuracy of the parameters after the recalibration step.

   To simplify the analyses, in this scenario, we will consider priors only on the faulty parameters, while keeping the others fixed at their estimates (given in θ_0). This is, similar to what mentioned in Section 2.3.3, an effective simplification of step 3 of running the analysis, where we update the parameter set in multiple iterations, starting off with ones that we are most uncertain about while keeping others fixed at their priors or point estimates, and then moving onto the next batch if the validity of the model is not still achieved. Plots for this setting are given in Figs. 5–7.

3. **Setting 3: Faulty model structure:** The model structure itself is wrong, which means that $M(x,\theta)$ does not correctly specify the reality, even if efforts are made to recalibrate the model in its current form with its given community-specific and global parameters. We create this scenario by producing a reality which is a distortion of the model outputs, that is, we define the alternative (true underlying) model structure as $(1+c)M_W$ instead of the presumed structure M_W, meaning we add $100c$% bias to the untransformed model outputs, where c is a constant. Thus, we can write $y_{R,k} = M(x_k, \theta_0) + D^2(x_k, \theta_0)$, where

$$D^2(x_k, \theta_0) = \operatorname{logit}((1+c)M_W(x_k, \theta_0)) - \operatorname{logit}(M_W(x_k, \theta_0))\,.$$

In this case, we should be able to falsify the model with running either of the steps 1 and 3 from Section 2.3.3. Plot for this setting is given in Fig. 8. The value of c chosen for this exercise is $c = 0.25$.

3.4. Results

The results from the above analyses are presented in Figs. 4–8 and in Table 1. Each plot, except Fig. 7, shows the posterior distribution of the discrepancy $D_{M_W}(x_k, \Theta)$ for communities $k = 1,\ldots,6$ under the different simulation scenarios. Fig. 7 shows the posterior distribution of the parameter σ_1, the rate of progression from infected state $I1$ to $I2$, with and without performing the recalibration step under setting 2 (Faulty prior on Θ). The analyses are repeated for multiple realizations of the observed data, resulting in different density curves for the discrepancy in each community for each such realization. In Figs. 4–8, 10 such curves are presented for each simulation scenario. We also calculate the posterior tail probability (as discussed in Section 2.3.4) for each scenario, aggregated over the different MCMC runs, and present the results in Table 1. Some of our observations can be summarized as the following:

3.4.1. Null case: both model and priors are accurate—Under the first simulation setting (the null scenario), the discrepancy for all six communities are centered around 0 both when $\mathscr{H}_\Theta$ is not updated, and also when we do update it (see Fig. 4). This is expected, since the discrepancy distribution should be centered around 0 when model assumptions are satisfied. However, some individual runs are not centered at 0 at all communities, even after updating — this might be because in those runs, the sampling errors for the communities in question are large.

Updating the prior, increased $\hat{p}_{TP}$ from 0.83 to 0.88 Table 1. Under H_0, p_{TP} should eventually converge 1, given an infinite number of samples. Under finite sample setting, the posterior mean will never be exactly at **0** due to both random fluctuations in the MCMC algorithm and the prior on λ_d, and hence the estimated values $\hat{p}_{TP}$ will always be less than 1. This provides a benchmark for how good we should expect a model to perform with respect to this metric, given that a ‘perfect’ model scores about 0.85 here.

3.4.2. Recalibrating faulty priors—Under the second simulation scenario, the prior for σ_1, the rate of progression from infected state $I1$ to $I2$, is misspecified downwards by a margin of 50%. We consider two types of faulty priors, (i) Flat — logitgauss prior with scale $p = 0.1$ (Fig. 7A), (ii) Narrow — logitgauss prior with scale $p = 0.01$ (Fig. 7B). Although both of these priors for σ_1 are shifted by the same amount, the flat prior assigns a substantial probability to the true value (black vertical line in 7 than the narrow prior, where it assigns no probability at all).

For both the narrow and flat prior, the model discrepancy is biased substantially away from zero in almost every community when the priors are not updated (as seen in Figs. 5 and 6). However the posterior distribution of D_M includes zero with a higher posterior tail probability in the flat prior than in the narrow prior. This is reflected in the moderately low p_{TP} of 0.63 for the flat prior, compared to 5.3×10^{-4} for the narrow prior.

When the prior of σ_1 is updated in the MCMC analysis, we can obtain a posterior distribution (for σ_1) centered around the true value only it is in the support for the faulty prior originally (Fig. 7). As a result, the posterior distribution of the discrepancy shifts towards 0 in the flat prior but not in the narrow prior (see Figs. 5 and 6), after the recalibration step is performed. After recalibration, the p_{TP} of the flat prior improves to 0.81, which is comparable to the null case. On the other hand, for the narrow prior, p_{TP} only improves to 0.088. Thus, even though the model structure is correct, whether we are able to recalibrate the model and verify its validity against the collected data can depend purely on the quantification of uncertainty in the stored value of its parameters, and whether the support of their stored stochastic information contain their true values.

3.4.3. Faulty model structure—

Under the third simulation setting, the underlying model structure itself is incorrect, producing biased model outputs, and as a result, the discrepancy is expected to be centered away from 0 even when updating the prior distributions $\mathscr{H}_\Theta$ for all parameters. Even though that is what we overwhelmingly see in Fig. 8, and that most simulations show a positive bias, one or two curves still appear to be centered around 0. So for those individual simulations, there are not any perceivable difference between the results and model predictions, even though the underlying model structure is faulty. Moreover, this mean shift is often less visible for some communities than others, for example in the case of community 3, which had the lowest prevalence in 2013. Since this scenario was created by inflating the model outputs by 25%, very little discrepancy was introduced in this community, and thus if we were to make a decision about whether the posterior distribution for the discrepancy is centered away from 0 or not, based on individual Bayesian credible intervals at 5% level of error, the hypothesis may be rejected for some communities, but not for others. In addition, for communities with very low prevalence, the sampling error is much higher, and can sometimes dominate the actual discrepancy in the outputs. Under faulty model scenario, $\hat{p}_{TP}$ values are low, whether we update the priors or not, but one thing that we notice from Table 1 is that updating the priors (recalibrating) always results in higher values of $\hat{p}_{TP}$.

4. Discussion

In this article, we outline a framework for evaluating the predictions of complex epidemiological models and describe experiments that can be used to test them. We propose assessing models by calculating the posterior distribution of the model discrepancy using a Bayesian framework. This allows for rapid identification of communities and/or subgroups for which the model performs poorly, and allows for an overall (or locally for each community) goodness of fit evaluation using the posterior tail probability. This methodology can then be systematically applied to update model priors to improve fit.

We apply this framework to a simple model of a heterosexual HIV epidemic, M_W, which was created to investigate how the proportion of early transmission affects the impact of ART on reducing HIV incidence. We test the model under various set of assumptions, including the null scenario when all the assumptions are satisfied, along with alternate scenarios when one or more of those assumptions fail to hold, and discuss the scope of recalibration of its parameters when the model is false.

One interesting finding is the existence of ‘uninformative communities’, that is, communities that do not provide information on model validity. In our case, community 3 acts in this way. It had very low prevalence initially, and also at each of the subsequent time points. Looking at the discrepancy plots for community 3 in Figs. 4–8, it becomes obvious that even if there is true discrepancy in the model output, due to a faulty model structure or because one or more of its parameter distributions are wrongly estimated, there is little or no visible sign of that in any of these plots, that is, most of the density plots can be seen to be centered around 0. If anything, the posterior distributions show a higher variability than that at other communities, hinting at the uninformativeness of this particular community.

The complexity of epidemiologic models creates a robustness that makes it quite challenging to falsify them. For example, in the posterior distribution plots for the discrepancy for the faulty model scenario (see Fig. 8), we see that even though most simulations show a positive bias, one or two of the curves do appear to be still centered around 0. So for those individual simulations, there are not any perceivable difference between the results and model predictions, even though the underlying model structure is faulty. Moreover, this mean shift is often less visible for some communities than others (like in the case of community 3), and hence if we were to make a decision about whether the posterior distribution for the discrepancy is centered away from 0 or not, based on individual Bayesian credible intervals at 5% level of error, the hypothesis may be rejected for some communities/ subgroups, but not for others. Although these individual inferences are crucial to figure out for which communities the model fails, it is also important to make inference on the overall strength of the model, that is, to aggregate the inferences over these different communities/ subgroups. The measure, p_{TP}, defined in Section 2.3.4 based on Mahalonobis distances, helps mitigate that issue somewhat. Delineating faulty model structure from insufficient calibration is a crucial aspect of model validation. This motivated our focus on these scenarios separately.

The methods developed here are for a fixed point of time at which model outputs are recorded and compared with external data. However, these methods can potentially be extended to incorporate comparison of model projections at multiple time points with real-time data as it becomes available over time. For this, one needs to also account for inter-person correlation and time series effects, and it might be worthwhile to pursue this extension in future research. Also note that the (prior) distribution for λ_d, the precision parameter for $\mathbf{D}_M$ was chosen in a data-driven manner, and the rest of the hyperparameters for the GASP prior on $\mathbf{D}_M$ were either fixed at pre-specified or data-driven values. It might be an interesting exercise to conduct a sensitivity analysis to explore the effect of this formulation on the analysis, for example by considering non-data-driven priors for λ_d with varying width, and see how that affects the results.

Note that there can be many different ways a model can have a faulty structure. A faulty structure for an ODE model means that mechanisms and processes which are essential for the population and transmission dynamics are not properly represented in the model. This can be due to multiple factors, including (i) not capturing essential confounders in the model, (ii) not accounting for human mobility patterns in the ODEs, (iii) unexpected events (for example a pandemic, intervention rollout, changes in standards of care) that might render the model structure outdated. Incorporating those factors in the model requires extensive structural changes which were outside the scope of this project. Instead, we assumed that whenever the model structure is wrong, the final effect will necessarily be seen in the predictions, which will be biased estimates (the bias coming from the faulty structure) of the reality. Hence, we decided to mimic the faulty structure by introducing bias in model outputs, without discussing the source of the bias, which can be due to any of the above reasons, or other sources.

Also note that from Section 2, our methods will continue to work for any model of the form $y = M(x, \theta)$, where $M(\cdot, \cdot)$ is essentially a function with arguments x and θ. Thus, the Bayesian validation framework will work for other types of models such as network models and stochastic agent-based models, as long as their parameter sets x and θ have concrete forms and are estimable (identifiable). However, one limitation the Bayesian validation framework is the fact that the MCMC analysis relies on multiple model runs over different instances of x and θ generated from the priors. ODE models are generally simple to run and computationally less expensive compared to other models (like agent-based models, network models), which might make the validation procedure extremely computationally burdensome, and therefore, infeasible to run in practice.

It is also worthwhile to mention that in the context of HIV transmission models, reduction in HIV incidence is a better choice as an outcome of interest for evaluation of the impact of real-world interventions. Although the HIV transmission model that we use in our simulation settings was designed to project HIV prevalence, our methods are applicable to modeling analyses where incidence is being estimated. In addition, we note that while incidence is a more desirable metric than prevalence, it is also much harder to estimate than prevalence.

Given a model structure and its parameter priors in their current states and the observed data, ‘optimal’ discrepancy is achieved through re-calibration of its parameters in the Bayesian validation framework but re-calibration is unlikely to reduce the discrepancy to 0. Rather, re-calibration would ensure the smallest possible discrepancy is achieved based on what we knew before (the model) and what we know now (the observed data). To achieve this ‘optimal’ discrepancy, our suggestion is to re-calibrate the parameters in steps, possibly one at a time, where the order can be determined by running sensitivity analyses to gauge the influence of each of these parameters on the final model outputs, and then updating the parameters in order of importance.

For parameters that have identifiability issues, model validation will have to endure some of the same challenges that model calibration face. For example, multimodality in a parameter distribution might indicate presence of latent groups or other structural issues in the model, which may or may not result in multimodality in the discrepancy distribution as well, all of which are important issues to consider in model validation, and we believe that the Bayesian validation framework can be used to diagnose these issues further. As far as the posterior tail probability is concerned, it should be noted that the posterior tail probability alone should not be the only aspect considered in making a decision on model validation, and other visual aspects like multimodality suggesting presence of latent groups or other issues should be taken into consideration as well. Multimodality in the discrepancy distribution may affect the posterior tail probability, for example in the scenario when the discrepancy distribution is bimodal, and the two modes of the discrepancy distribution lie on either side of 0, and the mean lies in between and closer to 0, we might obtain high p_{tp} indicating good model fit when clearly there are structural issues in the model. This is because p_{tp} in its current form is effective only when the discrepancy distribution is unimodal, as it defined around the mean of the discrepancy distribution assuming unimodality (see Section 2.3.4). So, when the discrepancy distribution is indeed multimodal, it will be more useful to redefine it in terms of the modes of the discrepancy distribution, and run multiple posterior tail probability analyses, one for each mode.

The HIV Modeling Consortium (www.hivmodelling.org) is a large network of mathematical modelers that aims to strengthen the use of models in decision making in HIV. In the past it has brought together models from different groups to quantify and characterize the extent to which different models predict different impacts of the same interventions (Eaton et al., 2015). The project has revealed a large amount of variation in model outputs, which has led to urgent questions being asked about whether models can be validated. To investigate this further, the Consortium developed a protocol for archiving models and their predictions in 2012, and invited researchers to submit their mathematical models for this exercise. With the methods described in this article, these models can be meaningfully tested and validated, and possibly even recalibrated for future use.

Acknowledgments

Sayan Dasgupta and James P. Hughes were supported by the National Institute of Health grant NIH-MH105857 for this work, while Mia R. Moore and Dobromir T. Dimitrov were supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health, United States through award UM1AI068617 (Statistical and Data Management Center: HIV Prevention Trials Network). Sayan Dasgupta will like to thank members of the Chen group at Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center for their many helpful suggestions.

Appendix A. Details of the modified Bayesian procedure

Here we present the details of the Bayesian analysis under certain modified conditions that we have adopted in our simulation setting. Here, we assume that a) we have high level estimates for the community-specific parameters in each community k, that is, $X_k;= x_{0,k}$, and b) λ_F is explicitly known through the design. Denote $\mathbf{x}_0 = \{x_{0,1},\ldots,x_{0,K}\}$.

The multivariate normal density for the collection of all field data $y_{F,1}, \ldots, y_{F.K}$ can now be denoted by $f\left(y_{F,1}, \ldots, y_{F,K} \mid \Theta, \mathbf{x}_0, \lambda_F(\mathbf{x}_0)\mathbf{D}_M, \lambda_d, \mu_d, \beta_d, \alpha_d\right)$, and the posterior density of the unknowns given the data y_F can be simplified (after removing the fixed quantities from the conditional distributions) as

$$\begin{aligned} &P(\Theta, D_M(x_{0,1}, \Theta), \ldots, D_M(x_{0,K}, \Theta), \lambda_d \mid y_{F,1}, \ldots, y_{F,K}, \mathbf{x}_0) \\ &\propto f(y_{F,1}, \ldots, y_{F,K} \mid \Theta, \mathbf{x}_0, D_M(x_{0,1}, \Theta), \ldots, D_M(x_{0,K}, \lambda_d)) \\ &\times P(\Theta, D_M(x_{0,1}, \Theta), \ldots, D_M(x_{0,K}, \Theta), \lambda_d \mid \mathbf{x}_0)\,. \end{aligned} \tag{9}$$

Appendix B. Metropolis Hastings algorithm

The Metropolis Hastings (MH) is an efficient method for sampling of the random variables (Θ,$\mathbf{D}_M$,λ_d) from their posterior distribution. We divide our parameters into three classes, $\Theta = \{\Theta_1, \Theta_2, \Theta_3\}$, where (i) Θ_1 are parameters that we want to fix at degenerate values (prior means)$\tilde{\theta}_1$, that is, $\Theta_1 : = \tilde{\theta}_1$, (ii) Θ_2 are parameters with prior distribution $\mathscr{H}_2(\tilde{\theta}_2)$ which we do not want to update, and (iii) Θ_3 are parameters with prior distribution $\mathscr{H}_3(\tilde{\theta}_3)$ which we want to update. The posterior distribution that we want to sample from is given as $P(\Theta_3, \mathbf{D}_M, \lambda_d \mid y_F, \mathbf{x}_0, \tilde{\theta}_1, \Theta_2)$.

At each step in the MH algorithm, we define a proposal distribution for each of the random variables whose joint posterior distribution is of interest to us. Thus, we assume we have proposal distributions q_{Θ_3}, $q_{\mathbf{D}_M}$, and q_{λ_d} for Θ_3, $\mathbf{D}_M$ and λ_d such that we can draw values $\theta_{3i} \sim q_{\Theta}(\,\cdot \mid \eta_0)$, $\mathbf{d}_{Mi} \sim q_{\mathbf{d}_M}(\,\cdot \mid \mathbf{d}_{M0})$, and $\lambda_{di} \sim q_{\lambda_d}(\,\cdot \mid \lambda_{d0})$ given some parameters η_0, $\mathbf{d}_{M0}$ and λ_{d0}. In the first step, we draw initial values $(\theta_2^{(0)}, \theta_3^{(0)}, \mathbf{d}_M^{(0)}, \lambda_d^{(0)})$ for (Θ_2,Θ_3,$\mathbf{D}_M$,λ_d) directly from the prior, that is,$(\theta_2^{(0)}, \theta_3^{(0)}, \mathbf{d}_M^{(0)}, \lambda_d^{(0)}) \sim \mathscr{H}_2(\tilde{\theta}_2)\mathscr{H}_3(\tilde{\theta}_3)P(\mathbf{D}_M, \lambda_d \mid \mathbf{x}_0) : = P(\Theta_2, \Theta_3, \mathbf{D}_M, \lambda_d \mid \mathbf{x}_0)$. Then the ith iteration of the algorithm is conducted in the following steps:-

1. Generate candidate proposals $(\theta_3^{(c)}, \mathbf{d}_M^{(c)}, \lambda_d^{(c)})$ as $\theta_3^{(c)} \sim q_{\Theta_3}(\,\cdot \mid \theta^{(i-1)})$, $\mathbf{d}_M^{(c)} \sim q_{\mathbf{D}_M}(\,\cdot \mid \mathbf{d}_M^{(i-1)})$, and $\lambda_d^{(c)} \sim q_{\lambda_d}(\,\cdot \mid \lambda_d^{(i-1)})$.

2. Generate $\theta_2^{(i)}$ as $\theta_2^{(i)} \sim \mathscr{H}_2(\tilde{\theta}_2)$.

3. Calculate the acceptance probability at the ith step, α_i, as Eq. (10) given in Box I.

4. Accept the candidate sample $(\theta_3^{(i)} : = \theta_3^{(c)}, \mathbf{d}_M^{(i)} : = \mathbf{d}_M^{(c)}, \lambda_d^{(i)} : = \lambda_d^{(c)})$ with probability α_i or reject it $(\theta_3^{(i)} : = \theta_3^{(i-1)}, \mathbf{d}_M^{(i)} : = \mathbf{d}_M^{(i-1)}, \lambda_d^{(i)} : = \lambda_d^{(i-1)})$ with probability $1 - \alpha_i$.

After a mandatory burn-in period, the resulting samples (a set of N draws) are drawn from the posterior distribution of Θ_3, λ_d, $\mathbf{D}_M$ (and resultantly ($\mathbf{x}_0$,Θ)).

Appendix C. Model equations for the HIV transmission model M_W

The Model diagram in Fig. 3 is described by a set of eight ordinary differential equations described below.

$$\frac{dS_1}{dt} = B\tau c - \mu_1 S_1$$

$$\frac{dS_2}{dt} = B(1-\tau)c - (\lambda_1 + \mu_1)S2$$

$$\frac{dS_3}{dt} = B(1-\tau)(1-c) - (\lambda_2 + \mu_1)S_3$$

$$\frac{dS_4}{dt} = B\tau(1-c) - \mu_1 S_4$$

$$\frac{dI_1}{dt} = \lambda_1 S_2 + \lambda_2 S_3 - (\mu_1 + \sigma_1)I_1$$

$$\frac{dI_2}{dt} = \sigma_1 I_1 - (1 + \mu_1 + \mu_2)I_2$$

$$\frac{dAI}{dt} = \alpha I_2 - (\mu_1 + \mu_3)AI$$

$$\frac{dNAI}{dt} = (1-\alpha)I_2 - (\mu_1 + \mu_2)NAI$$

where $S1$ denotes circumcised and susceptible men who are not at risk of infection, $S2$ denotes circumcised and susceptible men who are at risk of infection, $S3$ denotes uncircumsized and susceptible men who are at risk of infection, $S4$ denotes uncircumcised and susceptible men who are not at risk of infection, $I1$ denotes men in the first HIV infection state, $I2$ denotes individuals in the second (late stage) infection state, AI denotes infected men who will receive the antiretroviral therapy (as part of the intervention package) in 2015, and NAI denotes infected men who will never receive ART or other interventions.

The number of individuals entering the population is denoted by B, while the rate of individuals leaving the population is denoted by μ_1. The proportion of individuals entering the population who are not at risk is denoted by τ, and c is the proportion of those entering the population who are circumcised men. λ_1 and λ_2 describe the force of infection in circumcised and uncircumcised individuals respectively. σ_1 describes the rate of progression from infected state I_1 to I_2, and α is the proportion of individuals in state I_2 who will receive ART. μ_2 is the rate of death in late stage positive individuals (I_2) and μ_3 is the rate of death for individuals on ART treatment.

The force of infection is described by the following equations:

$$\lambda_1 = (1 - \varepsilon_C)\beta(t)\eta\frac{I_1 + I_2 + NAI + (1 - \varepsilon_A)AI}{S_1 + S_2 + S_3 + S_4 + I_1 + I_2 + AI + NAI}$$

$$\lambda_2 = \beta(t)\eta\frac{I_1 + I_2 + NAI + (1 - \varepsilon_A)AI}{S_1 + S_2 + S_3 + S_4 + I_1 + I_2 + AI + NAI}$$

where ε_C is the reduction in the risk of acquisition of HIV in circumcised men, $\beta(t)$ is the rate of transmission, η is used to modulate the force of infection to simulate a behavior change intervention, and ε_A is the reduction in the rate of transmission from individuals on treatment. The number of individuals entering the population is described by:

$$B = (-\mu_1 + \varepsilon)P$$

where ε represents the population growth rate and P represents the population size 15 years prior to the current time period. P was defined in this way to avoid differences in prevalence caused by interventions affecting the birth rate. The population at the year of the intervention is scaled to ensure that all regions have the same population size at the start of the intervention period.

There is also an additional parameter 'beta trend parameter' which is applied in all communities and is used to scale the rate of transmission β. This is separate from the other parameters so that it may be incorporated into uncertainty analyses. If time t is less than the time when the trend is turned on (t_{trend}), β is equal to the initial value specified, if time is greater than t_{trend}, β is modified using the beta trend parameter (ω).

$$\text{If } t \le t_{\text{trend}},$$

$$\beta(t) = \beta,$$

$$\text{If } t > t_{\text{trend}},$$

$$\beta(t) = \beta^{(1 - (\omega(t - (t_{trend} - 1))))}$$

where t_{trend} corresponds to the year 2015 and as such this modulation only occurs once the fitting has been completed (we fit the model to 2013 prevalence). Both β and ω are the same across all communities.

References

Adamson B, El-Sadr W, Dimitrov D, Gamble T, Beauchamp G, Carlson JJ, Garrison L Jr., Donnell D, 2019. The cost-effectiveness of financial incentives for viral suppression: HPTN 065 Study. Value Health 22 (2), 194–202. [PubMed: 30711064]

Bayarri MJ, Berger JO, Paulo R, Sacks J, Cafeo JA, Cavendish J, Lin CH, Tu J, 2007. A framework for validation of computer models. Technometrics 49 (2), 138–154.

Betancourt M, 2017. A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint, arXiv:1701.02434.

Boily MC, Masse B, Alsallaq R, Padian NS, Eaton JW, Vesga JF, Hallett TB, 2012. HIV treatment as prevention: considerations in the design, conduct, and analysis of cluster randomized controlled trials of combination HIV prevention. PLoS Med. 9 (7), e1001250. [PubMed: 22807657]

Case KK, Gomez GB, Hallett TB, 2018. The impact, cost and cost-effectiveness of oral pre-exposure prophylaxis in sub-Saharan Africa: a scoping review of modelling contributions and way forward. J. Int. AIDS Soc. 22, e25390.

Caswell H, 1976. The validation problem. In: Systems Analysis and Simulation in Ecology, Vol. 4. pp. 313–325.

Chatzilena A, Van Leeuwen E, Ratmann O, Baguelin M, Demiris N, 2019. Contemporary statistical inference for infectious disease models using Stan. Epidemics 29, 100367. [PubMed: 31591003]

Currin C, Mitchell T, Morris M, Ylvisaker D, 1991. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. J. Amer. Statist. Assoc. 86, 953–963.

de Montigny S, Adamson BJ, Mâsse BR, Garrison LP, Kublin JG, Gilbert PB, Dimitrov DT, 2018. Projected effectiveness and added value of HIV vaccination campaigns in South Africa: A modeling study. Sci. Rep. 8 (1), 1–12. [PubMed: 29311619]

Dimitrov D, Moore JR, Wood D, Mitchell KM, Li M, Hughes JP, Donnell DJ, Mannheimer S, Holtz TH, Grant RM, Boily MC, 2019. Predicted effectiveness of daily and non-daily PrEP for MSM based on sex and pill-taking patterns from HPTN 067/ADAPT. Clin. Infect. Dis. just accepted.

Dorrington R, Johnson L, Budlender D, 2005. ASSA2003 AIDS and Demographic Models: User Guide. Centre for Actuarial Research, University of Cape Town, for the AIDS Committee of the Actuarial Society of South Africa, pp. 1–45.

Eaton JW, Bacaër N, Bershteyn A, Cambiano V, Cori A, Dorrington RE, Fraser C, Gopalappa C, Hontelez JA, Johnson LF, Klein DJ, 2015. Assessment of epidemic projections using recent HIV survey data in South Africa: a validation analysis of ten mathematical models of HIV epidemiology in the antiretroviral therapy era. Lancet Glob. Health 3 (10), e598–e608. [PubMed: 26385301]

Eaton JW, Menzies NA, Stover J, Cambiano V, Chindelevitch L, Cori A, Hontelez JA, Humair S, Kerr CC, Klein DJ, Mishra S, 2014. Health benefits, costs, and cost-effectiveness of earlier eligibility for adult antiretroviral therapy and expanded treatment coverage: a combined analysis of 12 mathematical models. Lancet Glob. Health 2 (1), e23–e34. [PubMed: 25104632]

Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB, 2012. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–7. Med. Decis. Mak. 32 (5), 733–743.

Glynn PW, Iglehart DL, 1990. Simulation output analysis using standardized time series. Math. Oper. Res. 15 (1), 1–16.

Granich RM, Gilks CF, Dye C, De Cock KM, Williams BG, 2009. Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. Lancet 373 (9657), 48–57. [PubMed: 19038438]

Guttorp P, Walden A, 1987. On the evaluation of geophysical models. Geophys. J. Int. 91 (1), 201–210.

Hallett TB, Gregson S, Mugurungi O, Gonese E, Garnett GP, 2009. Assessing evidence for behaviour change affecting the course of HIV epidemics: A new mathematical modelling approach and application to data from Zimbabwe. Epidemics 1 (2), 108–117. [PubMed: 21352758]

Halloran ME, Auranen K, Baird S, Basta NE, Bellan S, Brookmeyer R, Cooper B, DeGruttola V, Hughes JP, Lessler J, Lofgren ET, Longini IM, Onnela JP, Ozler B, Seage G, Smith TA, Vespignani A, Vynnycky E, Lipsitch M, 2017. Simulations for Designing and Interpreting Intervention Trials in Infectious Diseases. Unpublished report.

Hayes RJ, Moulton LH, 2017. Cluster Randomised Trials. Chapman and Hall/CRC.

Johnson LF, Mossong J, Dorrington RE, Schomaker M, Hoffmann CJ, Keiser O, Fox MP, Wood R, Prozesky H, Giddy J, Garone DB, 2013. Life expectancies of South African adults starting antiretroviral treatment: collaborative analysis of cohort studies. PLoS Med. 10 (4), e1001418. [PubMed: 23585736]

Johnson LF, White PJ, 2011. A review of mathematical models of HIV/AIDS interventions and their implications for policy. Sex. Transm. Infect. 87 (7), 629–634. [PubMed: 21685191]

Kennedy MC, O'Hagan A, 2001. Bayesian calibration of computer models. J. R. Stat. Soc. Ser. B Stat. Methodol. 63, 425–450.

Kennedy MC, O'Hagan A, Higgins N, 2002. Bayesian analysis of computer code outputs. In: Quantitative Methods for Current Environmental Issues. Springer, London, pp. 227–243.

Kretzschmar ME, van der Loeff MFS, Birrell PJ, De Angelis D, Coutinho RA, 2013. Prospects of elimination of HIV with test-and-treat strategy. Proc. Natl. Acad. Sci. 110 (39), 15538–15543. [PubMed: 24009342]

Poole D, Raftery AE, 2000. Inference for deterministic simulation models: The bayesian melding approach. J. Amer. Statist. Assoc. 95 (452), 1244–1255.

Ramin M, Arhonditsis GB, 2013. Bayesian calibration of mathematical models: Optimization of model structure and examination of the role of process error covariance. Ecol. Inform. 18, 107–116.

Robert CP, Casella G, 2004. Monte Carlo optimization. In: Monte Carlo Statistical Methods. Springer, New York, pp. 157–204.

Robert CP, Changye W, 2020. Markov Chain Monte Carlo methods, a survey with some frequent misunderstandings. arXiv preprint, arXiv:2001.06249.

Sacks J, Welch WJ, Mitchell TJ, Wynn HP, 1989. Design and analysis of computer experiments. Statist. Sci. 409–423.

Sampson PD, Guttorp P, 1999. Operational evaluation of air quality models. In: Novartis Foundation Symposium. pp. 33–45.

Schwartlander B, Stover J, Hallett T, Atun R, Avila C, Gouws E, Bartos M, Ghys PD, Opuni M, Barr D, Alsallaq R, 2011. Towards an improved investment approach for an effective response to HIV/AIDS. Lancet 377 (9782), 2031–2041. [PubMed: 21641026]

Stanecki K, Garnett GP, Ghys PD, 2012. Developments in the field of HIV estimates: methods, parameters and trends.

Wakefield J, 2013. Bayesian and Frequentist Regression Methods. Springer Science & Business Media.

2013. Global Update on HIV Treatment 2013: Results, Impact and Opportunities, Vol. 124. World Health Organization.

Woods B, Rothery C, Anderson SJ, Eaton JW, Revill P, Hallett TB, Claxton K, 2018. Appraising the value of evidence generation activities: an HIV modelling study. BMJ Glob. Health 3 (6), e000488.

Box I.

$$\alpha_i = \alpha\left(\theta_3^{(c)}, \mathbf{d}_M^{(c)}, \lambda_d^{(c)} \mid \tilde{\theta}_1, \theta_2^{(i)}, \theta_2^{(i-1)}, \theta_3^{(i-1)}, \mathbf{d}_M^{(i-1)}, \lambda_d^{(i-1)}\right)$$

$$= \min \left\{ \frac{q_{\Theta_3}\left(\theta_3^{(c)} \mid \theta_3^{(i-1)}\right) q_{\mathbf{d}_M}\left(\mathbf{d}_M^{(c)} \mid \mathbf{d}_M^{(i-1)}\right) q_{\lambda_d}\left(\lambda_d^{(c)} \mid \lambda_d^{(i-1)}\right) P\left(\theta^{(c)}, \mathbf{d}_M^{(c)}, \lambda_d^{(c)} \mid y_F, \mathbf{x}_0, \tilde{\theta}_1, \theta_2^{(i)}\right)}{q_{\Theta_3}\left(\theta_3^{(i-1)} \mid \theta_3^{(c)}\right) q_{\mathbf{d}_M}\left(\mathbf{d}_M^{(i-1)} \mid \mathbf{d}_M^{(c)}\right) q_{\lambda_d}\left(\lambda_d^{(i-1)} \mid \lambda_d^{(c)}\right) P\left(\theta^{(i-1)}, \mathbf{d}_M^{(i-1)}, \lambda_d^{(i-1)} \mid y_F, \mathbf{x}_0, \tilde{\theta}_1, \theta_2^{(i-1)}\right)}, 1 \right\}. \tag{10}$$

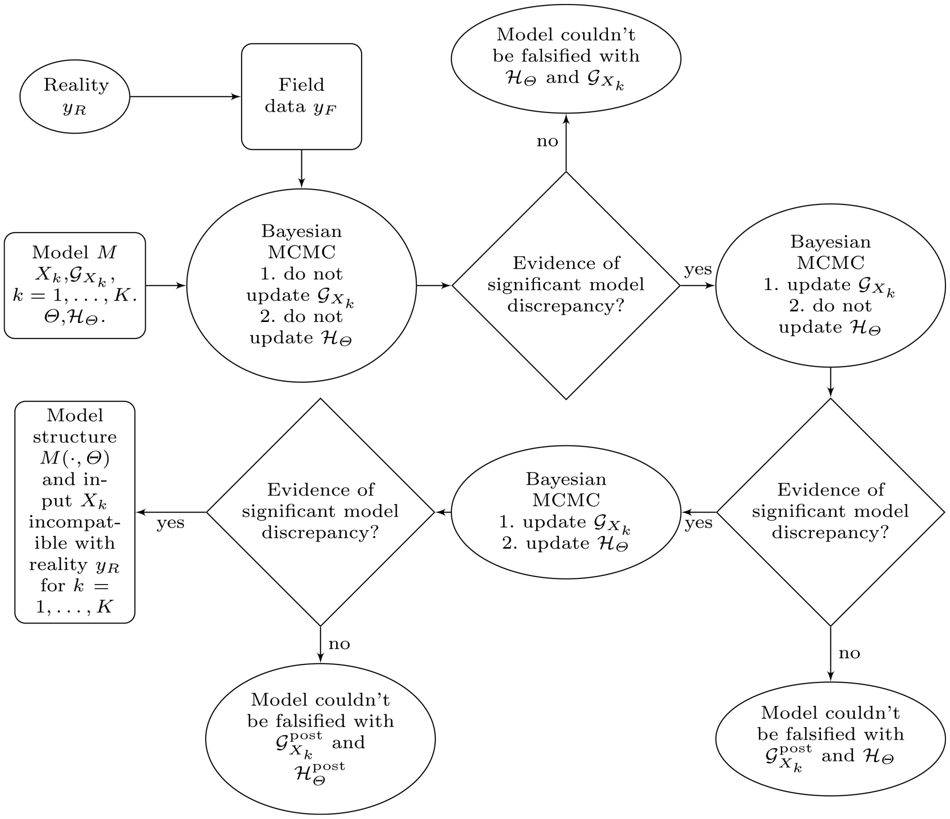
Reality y_R
Field data y_F
Model couldn't be falsified with $\mathcal{H}_\Theta$ and $\mathcal{G}_{X_k}$
no
Model M X_k,$\mathcal{G}_{X_k}$, $k = 1, \ldots, K$. Θ,$\mathcal{H}_\Theta$.
Bayesian MCMC 1. do not update $\mathcal{G}_{X_k}$ 2. do not update $\mathcal{H}_\Theta$
Evidence of significant model discrepancy?
yes
Bayesian MCMC 1. update $\mathcal{G}_{X_k}$ 2. do not update $\mathcal{H}_\Theta$
Model structure $M(\cdot, \Theta)$ and input X_k incompatible with reality y_R for $k = 1, \ldots, K$
yes
Evidence of significant model discrepancy?
Bayesian MCMC 1. update $\mathcal{G}_{X_k}$ 2. update $\mathcal{H}_\Theta$
yes
Evidence of significant model discrepancy?
no
no
Model couldn't be falsified with $\mathcal{G}^{\text{post}}_{X_k}$ and $\mathcal{H}^{\text{post}}_\Theta$
Model couldn't be falsified with $\mathcal{G}^{\text{post}}_{X_k}$ and $\mathcal{H}_\Theta$

Fig. 1.

One possible schematic of the Bayesian analysis for model validation.

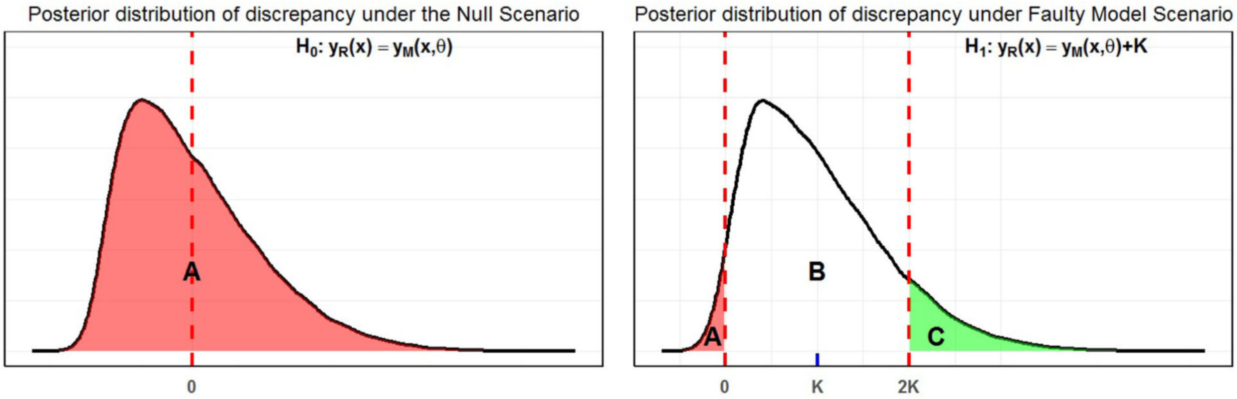

Fig. 2.

Posterior tail probabilities (p_{TP}), as denoted by the colored regions, under the Null and an hypothetical Faulty model structure scenario for discrepancy at a single community: (i) Under the Null, the mean of the discrepancy distribution is 0 itself, so any random sample from the discrepancy distribution will have Mahalonobis distance from its center greater than that between 0 and the center, and thus the posterior tail probability region will cover the entire discrepancy distribution, as shown by the shaded region A. (ii) Under the alternative, when the center of the discrepancy is at K units away from 0, the Mahalonobis distance between a randomly chosen sample and the center will be greater than that between the center and 0 if the sample belongs to either of the shaded regions A and C, and will be lesser than that between the center and 0 if it belongs to region B. Thus the posterior tail probability region will cover regions A and C. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

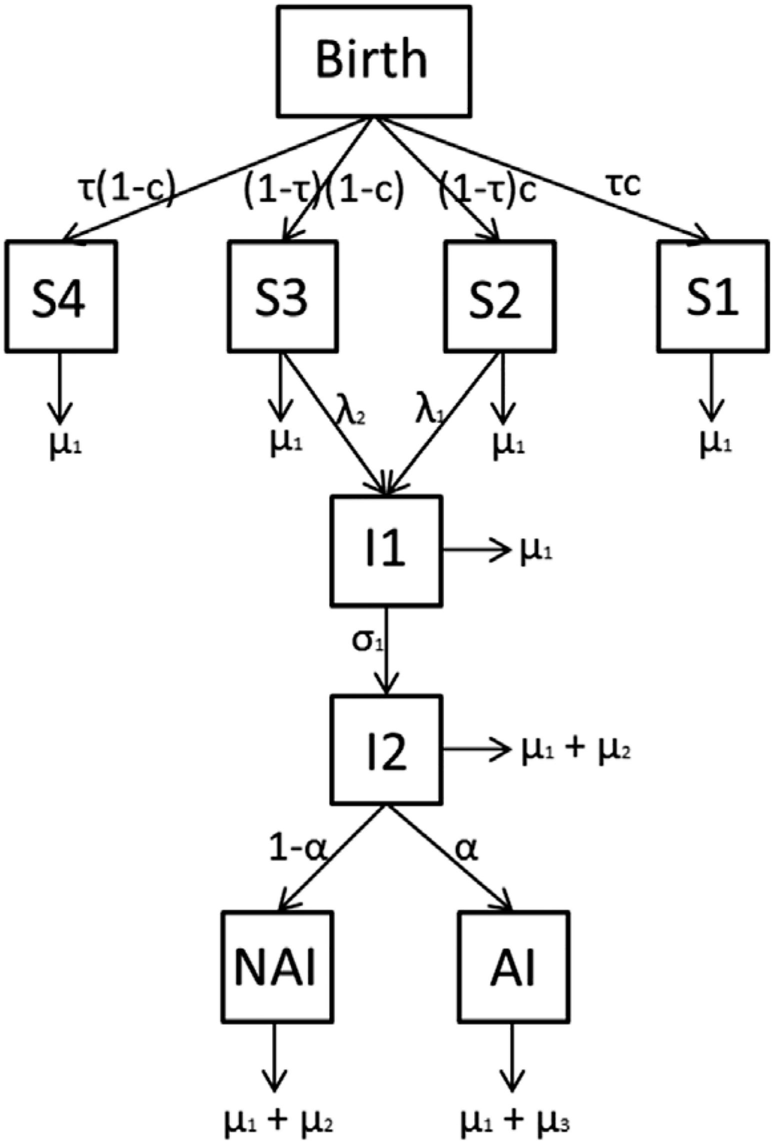

Fig. 3.

Model Diagram, *S*1 denotes circumcised and susceptible men who are not at risk of infection, *S*2 denotes circumcised and susceptible men who are at risk of infection, *S*3 denotes uncircumsized and susceptible men who are at risk of infection, *S*4 denotes uncircumcised and susceptible men who are not at risk of infection, *I*1 denotes men in the first HIV infection state, *I*2 denotes individuals in the second (late stage) infection state, *AI* denotes infected men who will receive the antiretroviral therapy (as part of the intervention package) in 2015, and *NAI* denotes infected men who will never receive ART or other interventions.

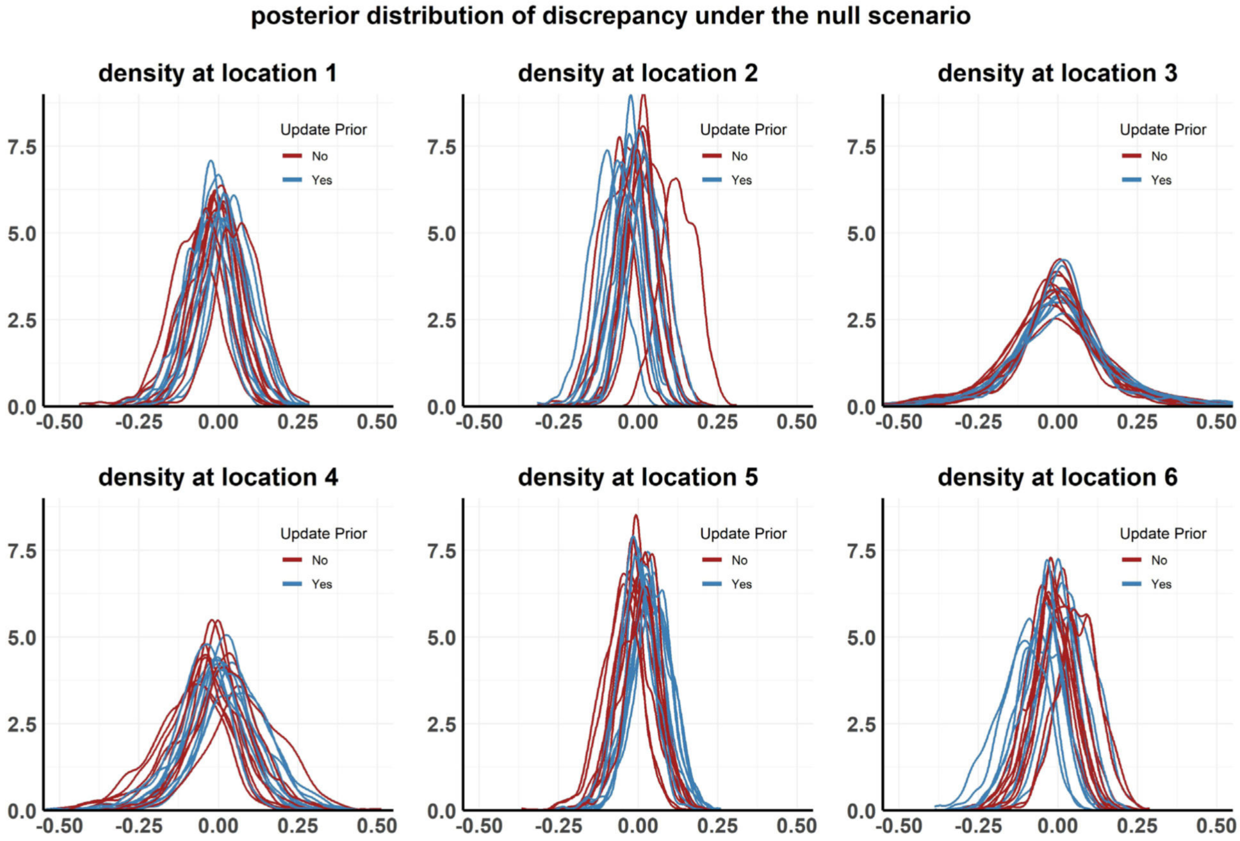

Fig. 4.

Setting 1: Posterior distribution of discrepancy in the null scenario for 10 MCMC chains, each run with a different realization of the observed data.

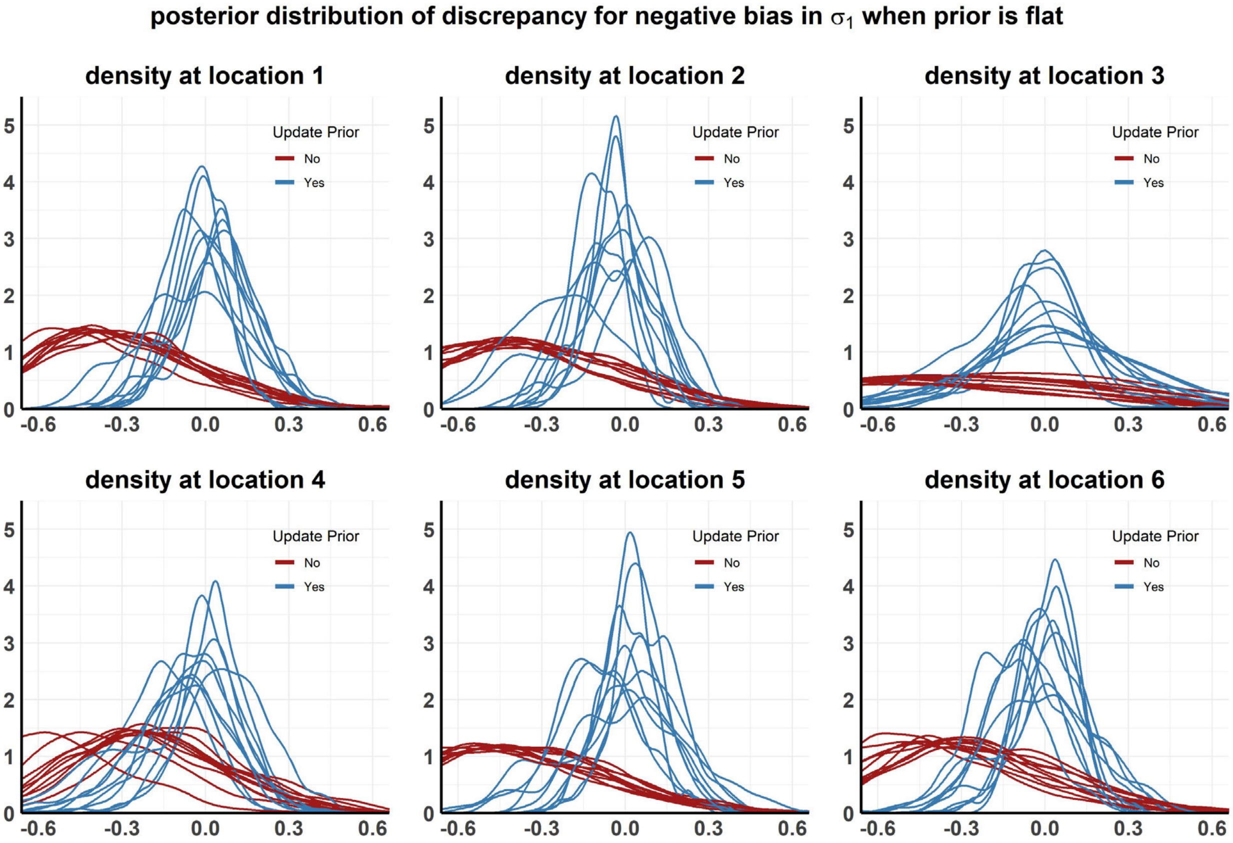

Fig. 5.

Setting 2: Posterior distribution of discrepancy when σ_1 is negatively biased and the prior width is **flat** under 10 MCMC chains, each run with a different realization of the observed data.

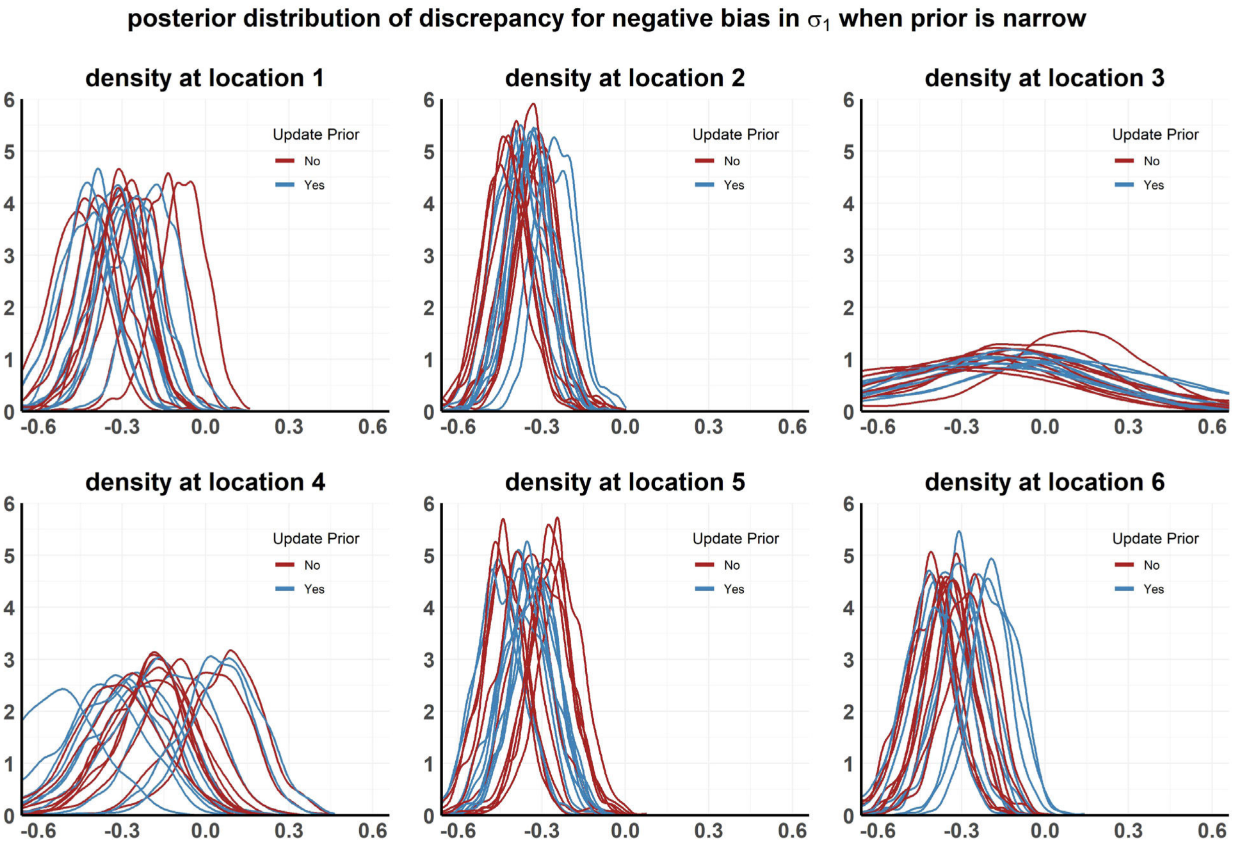

Fig. 6.

Setting 2: Posterior distribution of discrepancy when σ_1 is negatively biased and the prior width is **narrow** under 10 MCMC chains, each run with a different realization of the observed data.

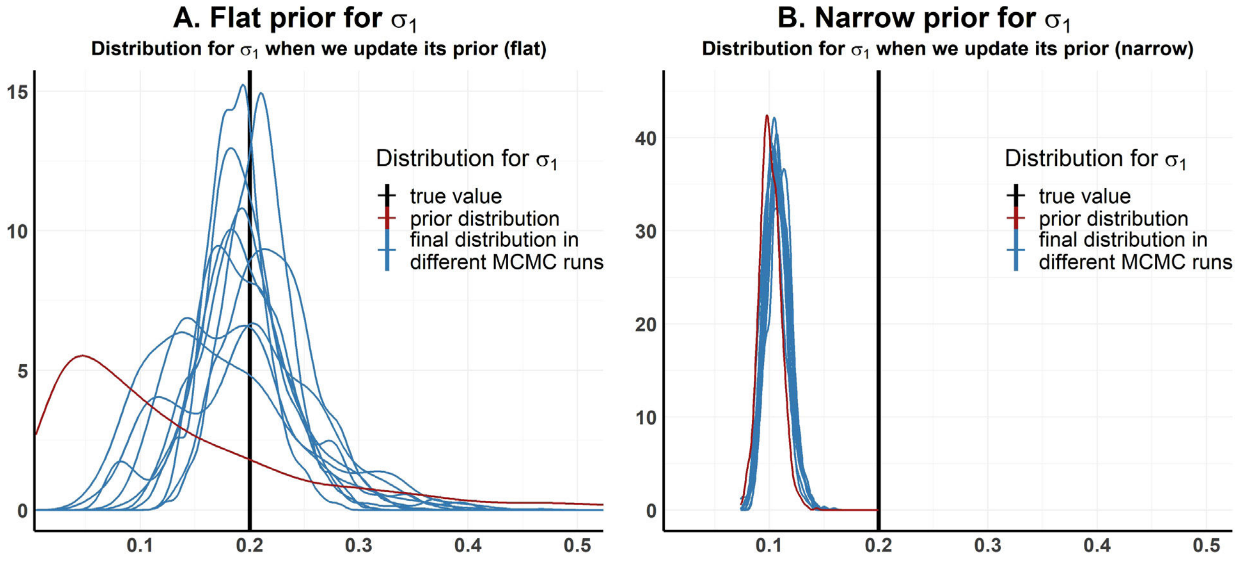

Fig. 7.

Setting 2: Posterior distribution of negatively biased σ_1 when we **update** $\mathscr{H}_{\Theta}$ and the prior width is **(A) flat** and **(B) narrow** under 10 MCMC chains, each run with a different realization of the observed data.

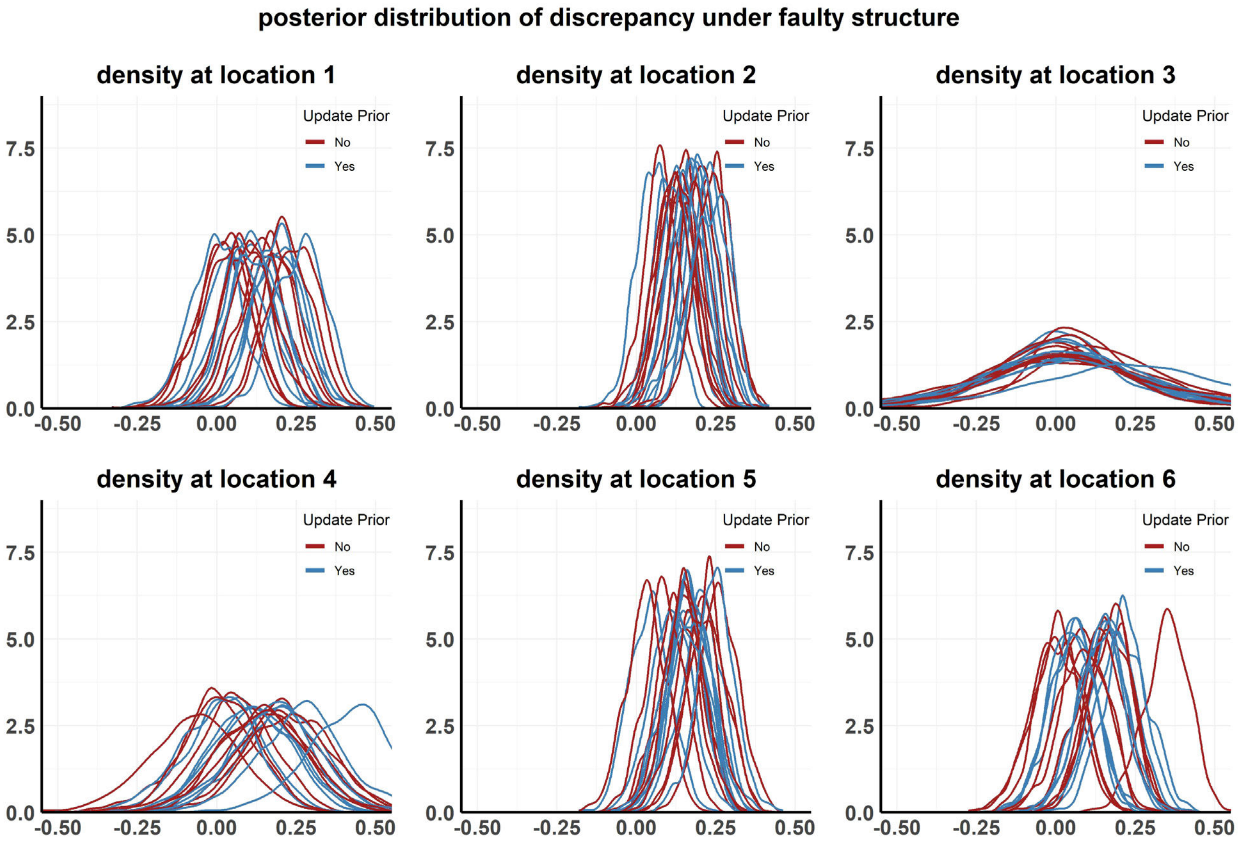

Fig. 8.

Setting 3: Posterior distribution of discrepancy under faulty model structure for 10 MCMC chains, each run with a different realization of the observed data.

Table 1

p_{TP} measures using Mahalonobis distance for (i) the Null scenario (ii) the Faulty model structure and (iii) Faulty prior information on Θ with different prior widths averaged over different MCMC runs.

Update H_{Θ}	**Null case**	**Recalibrating faulty prior**		**Faulty model structure**
		Flat prior	**Narrow prior**	
False	0.837	0.631	5.33e–04	0.148
True	0.886	0.810	0.088	0.212