When Is a Model Good Enough? Deriving the Expected Value of Model Improvement via Specifying Internal Model Discrepancies[*]

Mark Strong[†] and Jeremy E. Oakley[‡]

Abstract. A "law-driven" or "mechanistic" computer model is a representation of judgments about the functional relationship between one set of quantities (the model inputs) and another set of target quantities (the model outputs). We recognize that we can rarely define with certainty a "true" model for a particular problem. Building an "incorrect" model will result in an uncertain prediction error, which we denote "structural uncertainty." Structural uncertainty can be quantified within a Bayesian framework via the specification of a series of internal discrepancy terms, each representing at a subfunction level within the model the difference between the subfunction output and the true value of the intermediate parameter implied by the subfunction. By using value of information analysis we can then determine the expected value of learning the discrepancy terms, which we loosely interpret as an upper bound on the "expected value of model improvement." We illustrate the method using a case study model drawn from the health economics literature.

Key words. computer model, health economic model, model uncertainty, Bayesian decision theory, expected value of perfect information, Gaussian process

AMS subject classifications. 49Q12, 60G15, 62C99, 62P20, 91B74, 91G60

DOI. 10.1137/120889563

1. Introduction. When making predictions using computer models, two sources of uncertainty are uncertainty in model inputs and uncertainty in model structure. Methods for managing input uncertainty are well described (e.g., [17]), but methods for quantifying structural uncertainty are much less well developed. A "law-driven" or "mechanistic" model can be thought of as a representation of judgments about the relationship between one set of quantities, which forms the model input, and a set of target quantities represented by the model output. If we are uncertain what this true structural relationship is, then even if we were to run the model at its true inputs, there would be an uncertain "structural error" in the model prediction. We denote this uncertain error as "structural uncertainty." Note that we use the term "true" value of the input to mean that which we would estimate in some perfect study with infinite sample size, and "true" structural relationship to mean a (possibly nonunique) functional relationship that would result in the correct output given any set of true values of the inputs.

We denote the true unknown values of some vector of quantities that we wish to predict as Z (in general we will use uppercase Roman letters for unknowns and lowercase Roman letters

for realized values of unknown quantities). We represent our predictive computer model by the function $\boldsymbol{y} = f(\boldsymbol{x})$, where $\boldsymbol{x}$ is a vector of inputs and $\boldsymbol{y}$ is a vector of outputs in the same units as $\boldsymbol{Z}$. Note that there may be no closed-form expression for $f(\cdot)$, for example, if the model consists of a set of partial differential equations, or if the model is stochastic in some aspect. We may be uncertain about some or all of the components of $\boldsymbol{x}$ and write the true unknown values of the inputs as $\boldsymbol{X}$. If we are uncertain about the inputs, then this induces uncertainty in the outputs, and we write the vector of outputs (conditional on the model f) as $\boldsymbol{Y} = f(\boldsymbol{X})$. We may have additional uncertainty about $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$ if the model $f(\cdot)$ is stochastic, or if the model is computationally expensive such that we can evaluate $f(\cdot)$ only a small number of times. For the purposes of this paper we will assume that $f(\cdot)$ is deterministic and quick to evaluate, and therefore that there is no uncertainty about $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$.

This paper is motivated by the problem of managing model structure uncertainty in the context of health economic decision making. In the health economic setting, models are used to predict the future costs and health consequences of competing decision options in order to inform resource allocation decisions. Predictive models are used in much the same way as they are in the physical sciences, but with some important differences. The following features are typical:

- We wish to predict a (possibly vector) quantity $\boldsymbol{Z}$ using a computer model $f(\boldsymbol{X})$.
- We will use $\boldsymbol{Z}$ to inform a resource allocation decision.
- There is only one model.
- There are no observations on the target quantity $\boldsymbol{Z}$.
- The model inputs $\boldsymbol{X}$ are observable.
- The model is "simple" enough that we can "decompose" the model into subfunctions.

There are a number of approaches to managing uncertainty in model structure. The problem has been addressed from a statistical perspective, and here a key concept is that of *model averaging*. In this approach the predictions or probability statements of a number of plausible models are averaged, with weights based either on some measure of model adequacy or some measure of the probability that the model is true (key references are Bernardo and Smith [1], Draper [5], Kass and Raftery [10], and Kadane and Lazar [9]).

A somewhat different approach is described in the computer models literature. Here, attention is focused on the model *discrepancy*, $\boldsymbol{\epsilon} = \boldsymbol{Z} - f(\boldsymbol{X})$, the difference between the output of a model evaluated at the true inputs and the true target value. Observations (typically partial and/or noisy observations) on the target quantity $\boldsymbol{Z}$ are combined with the outputs of the computer model in order to update beliefs about $\boldsymbol{\epsilon}$, and hence about $\boldsymbol{Z}$. This process is known as *model calibration*, and important papers demonstrating methods include Kennedy and O'Hagan [11], Craig et al. [4], Higdon et al. [8], and Goldstein and Rougier [7]. Because $\boldsymbol{\epsilon}$ here relates to the difference between the model *output* and reality, we refer to this as the *external* discrepancy.

Neither of these approaches is sufficient in our context as outlined above. Model averaging methods require more than one model, and we typically have only a single model. Furthermore, even if we were to have multiple models, model averaging would still be difficult because we have no observations on the output quantity and therefore no way of weighting models based on either likelihood or predictive ability.

In contrast, calibration-based methods can be used in the single model context, but again

they rely on observations on the model output quantity. In health economics these observations are absent, and indeed models are employed primarily to make predictions in the absence of data that relate directly to the quantity of interest. Where there are no observations on Z we could, of course, elicit probability statements about the external model discrepancy ϵ, but judgments are likely to be crude.

Motivated by the problem of quantifying structural uncertainty in the health economic context, we outlined in a previous paper a novel approach based on an analysis of discrepancies *within* the model (which we call "internal" discrepancies) [18]. In the absence of data to inform discrepancies we are reliant on expert judgment in our management of model uncertainty, and our key proposition is that judgment about internal discrepancy is easier than judgment about external discrepancy. The approach in [18] allows the expert (which in many cases would be the modeler) to make judgments that can then be used to determine whether the model structure uncertainty has any impact on decision uncertainty, and if so, to guide further model development.

In [18] we illustrated the internal discrepancy method with a simple decision tree model, and our focus was on quantifying prediction uncertainty. In this paper we test the internal discrepancy method in a more complex type of health economic evaluation model, the Markov model, and our focus is on decision uncertainty rather than prediction uncertainty. The paper is organized as follows. In section 2 we describe the internal discrepancy method for managing model structure uncertainty. In section 3 we briefly review existing methods. In section 4 we introduce a case study based on a model designed to predict the costs and health effects of two competing treatment options for HIV/AIDS. In section 5 we apply the discrepancy analysis method to the case study model in three scenarios that represent plausible sets of assumptions regarding the structural error. In section 6 we present the results of a decision theoretic sensitivity analysis for each scenario. In a final section we discuss limitations and draw conclusions.

2. The internal discrepancy analysis method. The internal discrepancy method consists of four steps: model decomposition, identification of discrepancies, specification of beliefs about discrepancies, and sensitivity analysis. We outline each in turn before going on to discuss the management of complex "structured" discrepancies.

2.1. Model decomposition. First, the model is decomposed into a series of subfunctions, the outputs of which can be thought of as "intermediate" model parameters. The decomposition is chosen such that these intermediate parameters are potentially observable in the real world. To illustrate, Figure 1(a) shows a hypothetical model with ten inputs, $Y = f(X_1, \ldots, X_{10})$, that aims to predict a quantity Z. The model has been decomposed into a series of subfunctions, for example, $Y_1 = f_1(X_1, X_2, X_3)$ and $Y_2 = f_2(X_4, X_5)$, revealing a set of six "intermediate" parameters $Y_1, \ldots, Y_6$ that have "true" values $Z_1, \ldots, Z_6$.

Unless a model is trivially small there will be no unique decomposition. However, we expect that in most cases there will be a "natural" decomposition that reflects the manner in which the model has been constructed using smaller subunits.

2.2. Identification of discrepancies. For each subfunction we then ask the question, If we knew true values of the inputs to this subfunction, would we believe with certainty that the
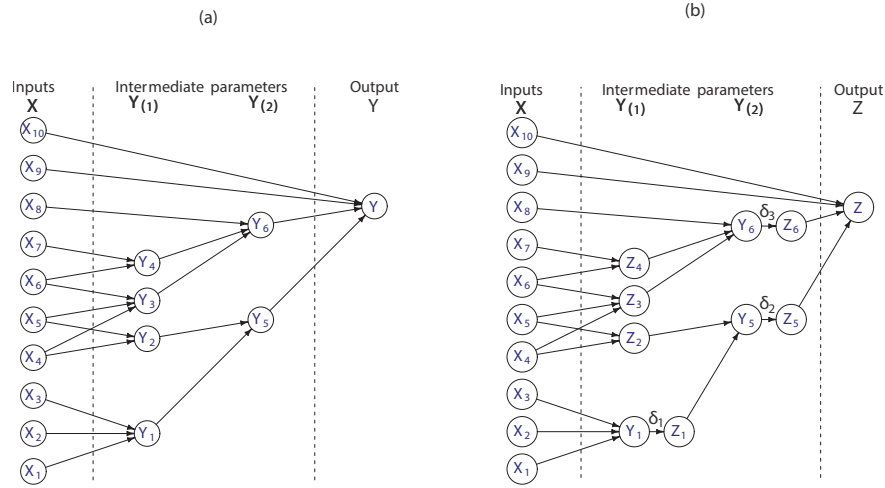
Figure 1. (a) *Hypothetical model with ten inputs and one output, decomposed to reveal six intermediate parameters.* (b) *We suppose that there is structural error in the subfunctions that result in Y_1, Y_5, and Y_6. Three discrepancy terms are added to correct the error, i.e., $Z_1 = Y_1 + \delta_1$, $Z_5 = Y_5 + \delta_2$, and $Z_6 = Y_6 + \delta_3$.*

output of the subfunction would equal the true value of the corresponding intermediate model parameter? Where we judge there to be potential structural error we introduce a discrepancy term δ_i. This allows us to write the *internal* model discrepancy equation

$$(2.1) \qquad\qquad \boldsymbol{Z} = f^*(\boldsymbol{X}, \delta_1, \ldots, \delta_I),$$

where $f^*(\cdot)$ is the functional form of an "augmented" model that includes internal discrepancy terms δ_i for the $i = 1, \ldots, I$ subfunctions in which we judge there to be error. In our hypothetical example we judge there to be structural error in three of the subfunctions. Figure 1(b) shows the incorporation of three uncertain internal discrepancy terms added to the model to correct the structural error.

2.3. Specification of beliefs about discrepancies. We express our beliefs about the size (and direction) of the errors via the (joint) distribution, $p(\boldsymbol{X}, \delta_1, \ldots, \delta_I)$, over internal discrepancies and inputs $\boldsymbol{X}$. We assume that we do not have direct observations on the intermediate parameters Z_i, and therefore we must use prior judgment to specify $p(\boldsymbol{X}, \delta_1, \ldots, \delta_I)$. However, if data *were* available on the intermediate parameters Z_i, then we could adopt the Kennedy and O'Hagan framework [11] and update beliefs about Z_i, and hence the target quantity $\boldsymbol{Z}$.

The key idea is that in some applications it is easier to make judgments about the internal discrepancies, $\boldsymbol{\delta} = \delta_1, \ldots, \delta_I$, than it is to make judgments about the external discrepancy, $\boldsymbol{\epsilon}$. The external discrepancy results from inadequacies throughout the whole model, and making judgments about the combined effect of the (possibly many) inadequacies is likely to be difficult. In contrast, the individual internal discrepancies relate only to small subfunctions, and the impact of deficiencies is expected to be easier to predict.

We recognize, however, that even where subfunctions are very simple, it may be difficult to make judgments about the internal discrepancies. We expect, therefore, that judgments are likely to be crude, but for the purposes of the sensitivity analysis all we require is that

the expression of uncertainty be "generous." By this we mean that the distribution chosen for δ_i has good coverage (say 99%) over all plausible values of δ_i that could arise from the subfunction inadequacy. So, we might use questions such as "could the true value of this parameter be twice as large as the modelled value, or ten times as large, or one hundred times as large?" to help to identify the range of values that are plausible. It is helpful to think of this process as defining the range of plausible scenarios, rather than as a formal elicitation exercise.

2.4. Sensitivity analysis. Finally, we determine the sensitivity of the model output (or a decision based on the model output) to the internal discrepancies. On this last point, we note that an important benefit arises from an analysis of internal discrepancy versus an analysis of external discrepancy. By determining the sensitivity of the model output to the internal discrepancies we gain some indication of the relative importance of the structural uncertainty within each model "subunit." We showed in [18] how such information can be used to guide further model improvement in those parts of the model where structural uncertainty has an impact on output uncertainty. More generally, if the output is insensitive to *all* the internal discrepancies, then this offers reassurance that our current model is "good enough." Conversely, if the output is highly sensitive to the discrepancies, then it may be advisable to consider more carefully the model structure in order to reduce structural uncertainty, or accept that there is considerable prediction uncertainty arising from the uncertain model structure.

The decision theoretic approach to determining sensitivity asks the question, What effect does changing the values of a group of inputs of interest, $\boldsymbol{X}_i$, have on the *decision* that will be made given the model output? We denote the utility of decision d, assuming model output $\boldsymbol{Y} = f(\boldsymbol{X})$, as $U\{d, f(\boldsymbol{X})\}$. The expected utility of our "baseline" optimum decision is

$$(2.2) \qquad \max_d \mathbb{E}_{\boldsymbol{X}}[U\{d, f(\boldsymbol{X})\}].$$

If we were to learn some group of inputs, $\boldsymbol{X}_i = \boldsymbol{x}_i$, then our maximized expected utility would be

$$(2.3) \qquad \max_d \mathbb{E}_{\boldsymbol{X}_{-i}|\boldsymbol{x}_i}[U\{d, f(\boldsymbol{x}_i, \boldsymbol{X}_{-i})\}],$$

where $\boldsymbol{X}_{-i}$ is the set of inputs not in $\boldsymbol{X}_i$. The maximized expected utility (2.3) itself is uncertain because $\boldsymbol{X}_i$ is uncertain, and it has expectation

$$(2.4) \qquad \mathbb{E}_{\boldsymbol{X}_i}(\max_d \mathbb{E}_{\boldsymbol{X}_{-i}|\boldsymbol{X}_i}[U\{d, f(\boldsymbol{X}_i, \boldsymbol{X}_{-i})\}]).$$

The expected gain in utility on learning $\boldsymbol{X}_i$ is the difference between (2.4) and the utility at baseline,

$$(2.5) \qquad \text{EVPI}(\boldsymbol{X}_i) = \mathbb{E}_{\boldsymbol{X}_i}(\max_d \mathbb{E}_{\boldsymbol{X}_{-i}|\boldsymbol{X}_i}[U\{d, f(\boldsymbol{X}_i, \boldsymbol{X}_{-i})\}]) - \max_d \mathbb{E}_{\boldsymbol{X}}[U\{d, f(\boldsymbol{X})\}].$$

This quantity is called the *expected value of perfect information* (EVPI) for $\boldsymbol{X}_i$ (or sometimes the "partial" EVPI, to reflect that we are conditioning on only a subset of the inputs) [16]. The overall EVPI for *all* inputs is given by

$$(2.6) \qquad \text{EVPI}(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{X}}[\max_d U\{d, f(\boldsymbol{X})\}] - \max_d \mathbb{E}_{\boldsymbol{X}}[U\{d, f(\boldsymbol{X})\}].$$

Given the augmented model f^*, we can determine the value of learning discrepancy terms,

$$(2.7) \qquad \mathrm{EVPI}(\boldsymbol{\delta}) = \mathbb{E}_{\boldsymbol{\delta}}(\max_d \mathbb{E}_{\boldsymbol{X}|\boldsymbol{\delta}}[U\{d, f^*(\boldsymbol{X}, \boldsymbol{\delta})\}]) - \max_d \mathbb{E}_{\boldsymbol{X}, \boldsymbol{\delta}}[U\{d, f^*(\boldsymbol{X}, \boldsymbol{\delta})\}],$$

relative to the value of learning the inputs,

$$(2.8) \qquad \mathrm{EVPI}(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{X}}(\max_d \mathbb{E}_{\boldsymbol{\delta}|\boldsymbol{X}}[U\{d, f^*(\boldsymbol{X}, \boldsymbol{\delta})\}]) - \max_d \mathbb{E}_{\boldsymbol{X}, \boldsymbol{\delta}}[U\{d, f^*(\boldsymbol{X}, \boldsymbol{\delta})\}].$$

This will give us some indication of the importance of the discrepancies in driving decision uncertainty. If $\mathrm{EVPI}(\boldsymbol{\delta})$ is large compared with $\mathrm{EVPI}(\boldsymbol{X})$, then we conclude that structural uncertainty is important. Partial EVPI analysis of each δ_i can then be used to identify those parts of the model in which structural uncertainty has an important impact on decision uncertainty.

We loosely interpret the expected value of learning the vector of discrepancy terms, $\mathrm{EVPI}(\boldsymbol{\delta})$, as the *expected value of model improvement* (EVMI). However, because model improvement is likely to involve the addition of new *uncertain* input parameters, $\mathrm{EVPI}(\boldsymbol{\delta})$ will provide an upper bound for the EVMI, rather than the EVMI itself. If the $\mathrm{EVPI}(\boldsymbol{\delta})$ is small, this offers us some reassurance that the model is good enough for the decision, whereas if it is large, we know our uncertainty about the model structure is resulting in decision uncertainty. In the latter case improving the model may be worthwhile, but this will depend on the degree of decision uncertainty induced by any newly introduced uncertain inputs.

Although the discrepancy terms $\boldsymbol{\delta}$ are in one sense just another set of inputs in the augmented model function $f^*(\boldsymbol{X}, \boldsymbol{\delta})$, there are two important differences between the treatment of $\boldsymbol{\delta}$ and $\boldsymbol{X}$. First, we learn about $\boldsymbol{X}$ directly from data, whereas we learn about $\boldsymbol{\delta}$ through expert elicitation. Second, decision uncertainty that results from uncertainty about $\boldsymbol{X}$ leads us to seek more data (i.e., to conduct further studies), whereas decision uncertainty that results from uncertainty about $\boldsymbol{\delta}$ leads us (at least in the first instance) to review model structure.

2.5. Managing complex "structured" discrepancies. So far we have assumed that internal discrepancies are scalar terms that relate to separate, distinct subfunctions within the model. In some models we may also wish to introduce discrepancies to the outputs of some set of coupled subfunctions. A Markov model is a good example. If we believe that a simple time-homogeneous Markov process does not reflect reality, then we may want to introduce discrepancies at each time step. Discrepancies in this context will be expected to have a correlation structure that reflects the nature of the process that is being modelled. One of the key aims of this paper is to explore how we can specify judgments about such a set of highly correlated discrepancies.

In most models we would expect there to be both "simple" discrepancies, judgments about which can be specified directly, and more complex "structured" groups of discrepancies that are specified through the hyperpriors of some form of flexible model. In our case study that follows we have chosen to concentrate our analysis on the latter.

3. Case study.

3.1. Description of model. Our case study is based on a four-state Markov model first described in [3] and subsequently used for illustrative purposes in [6] and [2]. The purpose of

the model is to predict costs and health outcomes (in life years) under two decision options, zidovudine monotherapy versus zidovudine plus lamivudine combination therapy, in people with HIV. We assume that the decision maker wishes to choose the therapy that maximizes expected utility, where utility for decision option d is defined as the monetary net benefit

$$(3.1) \qquad\qquad NB_d = \lambda E_d - C_d.$$

The terms E_d and C_d are, respectively, the total health outputs and total costs associated with decision d, and correspond to $\boldsymbol{Z}$ in the general notation in the introduction. The term λ is the decision maker's value in monetary units of one year of life. Note that NB_d in this case is scalar and is in monetary units. We do not believe that our case study model is the true model, and therefore the predicted costs and effects, which we denote by $\widehat{E}_d$ and $\widehat{C}_d$, do not necessarily equal their true values. The terms $\widehat{E}_d$ and $\widehat{C}_d$ correspond to $\boldsymbol{Y}$ in the general notation in the introduction.

Computer models in health economic analysis tend to be relatively simple compared with those used in the physical sciences. Typically, a notional population of individuals is modelled. For each decision option d, each individual in the population exists at time point $t \in (t_0, t_1)$ in one of a set of $n = 1, \ldots, N$ mutually exclusive and exhaustive health states. Each state is associated with health care costs C_{dnt} and health outputs E_{dnt} at each time point t. The proportion of the population that exists in state n at time t under decision option d is denoted by π_{dnt}, and we can therefore think of the vector $\boldsymbol{\pi}_{dt} = \{\pi_{d1t}, \ldots, \pi_{dNt}\}'$ as the "state vector" for decision d. Alternatively, π_{dnt} can be thought of as the probability that a single individual exists in health state n at time t given decision option d. Summing over the states and over time (in discrete units) gives us the total costs under decision d,

$$(3.2) \qquad\qquad \widehat{C}_d = \sum_{t=t_0}^{t_1} \sum_{n=1}^{N} \pi_{dnt} C_{dnt},$$

and total health output,

$$(3.3) \qquad\qquad \widehat{E}_d = \sum_{t=t_0}^{t_1} \sum_{n=1}^{N} \pi_{dnt} E_{dnt}.$$

In the simplest form of the Markov model, as implemented in this paper, transitions from one health state to another are assumed to follow a time-homogeneous Markov process (i.e., transition probabilities depend only on the current state and remain fixed for all time steps). Four health states are defined in the case study model with allowable transitions between states shown in Figure 2. The transition matrix for the monotherapy ($d = 1$) option is given by

$$(3.4) \qquad\qquad \boldsymbol{M}_1 = \begin{pmatrix} p_{111} & p_{112} & p_{113} & p_{114} \\ 0 & p_{122} & p_{123} & p_{124} \\ 0 & 0 & p_{133} & p_{134} \\ 0 & 0 & 0 & 1 \end{pmatrix},$$
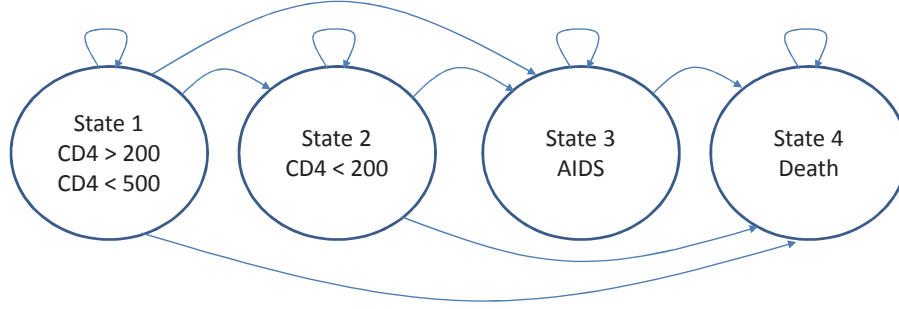
Figure 2. *Structure of the case study Markov model.*

where p_{1xy} denotes the probability of transition from state x to state y under decision option 1. To obtain $\boldsymbol{M}_2$, the transition matrix for the combination therapy decision option, the matrix $\boldsymbol{M}_1$ is modified by the incorporation of a parameter α, which quantifies the relative risk of disease progression in the combination therapy group versus the monotherapy group. This results in

$$(3.5) \qquad \boldsymbol{M}_2 = \begin{pmatrix} 1 - \alpha(p_{112} + p_{113} + p_{114}) & \alpha p_{112} & \alpha p_{113} & \alpha p_{114} \\ 0 & 1 - \alpha(p_{123} + p_{124}) & \alpha p_{123} & \alpha p_{124} \\ 0 & 0 & 1 - \alpha p_{133} & \alpha p_{134} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

We know that in our case study $0 < \alpha \leq 1$, and therefore $\boldsymbol{M}_2$ is a proper transition matrix. Given $\boldsymbol{M}_d$ and $\boldsymbol{\pi}_{d,t-1}$, we generate $\boldsymbol{\pi}_{dt}$ via

$$(3.6) \qquad \boldsymbol{\pi}'_{dt} = \boldsymbol{\pi}'_{d,t-1}\boldsymbol{M}_d,$$

and we can therefore express $\boldsymbol{\pi}_{dt}$, in terms of the state vector at time step 0, as

$$(3.7) \qquad \boldsymbol{\pi}'_{dt} = \boldsymbol{\pi}'_{d0}\boldsymbol{M}_d^t,$$

where $\boldsymbol{M}_d^t = \prod_{l=1}^t \boldsymbol{M}_d$. For our case study time is measured in discrete steps of 1 year and $t = 0, \ldots, 20$, where $t = 0$ is the time of treatment initiation and $t = 20$ is the "time horizon" of the analysis.

We denote the vector of costs for all health states at time step t under decision d as $\boldsymbol{C}_{dt} = (C_{d1t}, \ldots, C_{d4t})'$, and the vector of health effects as $\boldsymbol{E}_{dt} = (E_{d1t}, \ldots, E_{d4t})'$. Our overall model for the net monetary benefit associated with decision option d is therefore

$$(3.8) \qquad \widehat{NB}_d = \lambda \widehat{E}_d^{tot} - \widehat{C}_d^{tot} = \lambda \sum_{t=0}^{20} \boldsymbol{\pi}'_{d0}\boldsymbol{M}_d^t\boldsymbol{E}_{dt} - \sum_{t=0}^{20} \boldsymbol{\pi}'_{d0}\boldsymbol{M}_d^t\boldsymbol{C}_{dt}.$$

The quantities π_{dnt}, C_{dnt}, and E_{dnt} and the components of $\boldsymbol{M}_2$ can be thought of as "intermediate parameters," akin to the intermediate parameters in the hypothetical model shown in Figure 1.

$(p_{111}, p_{112}, p_{113}, p_{114}) \sim$ Dirichlet $(1251, 350, 115, 14)$
$(p_{121}, p_{122}, p_{123}, p_{124}) \sim$ Dirichlet $(0, 731, 512, 15)$
$(p_{131}, p_{132}, p_{133}, p_{134}) \sim$ Dirichlet $(0, 0, 1312, 437)$
$(p_{141}, p_{142}, p_{143}, p_{144}) = (0,\ 0,\ 0,\ 1)$

Table 2
Cost and relative risk distributions.

Label	Description	Distribution	Mean	SD
cc_1	Undiscounted care costs of one time step in state 1 (£)	Normal	2756	400
cc_2	Undiscounted care costs of one time step in state 2 (£)	Normal	3052	437
cc_3	Undiscounted care costs of one time step in state 3 (£)	Normal	9007	1449
α	Treatment effect (combi-therapy versus monotherapy)	Log-normal	$\log(0.509)$	0.05

In the context of health economic evaluation it is usual to define the utility of $\widehat{NB}_d$ as equal to $\widehat{NB}_d$ itself. Thus, the expected utility of the optimum decision is given by

$$(3.9) \qquad\qquad \max_d \mathbb{E}[U\{d, f(\boldsymbol{X})\}] = \max_d \mathbb{E}(\widehat{NB}_d).$$

3.2. Input parameter values. Transition probabilities, costs, and the treatment effect parameter are all considered uncertain in the base case model, with distributions shown in Tables 1 and 2. Input parameter estimates are all derived from primary studies (for further details, see references in [3]). Here we make the important point that the role of data in this case study is *solely* to inform the input parameter values. This is typical in health economic applications because observations on the model output quantity rarely exist.

Drug treatment costs are $c_Z = £2{,}278$ and $c_L = £2{,}087$ for zidovudine and lamivudine, respectively, and are considered known with certainty. Costs and effects accruing in the future (i.e., at time steps $t > 0$) are discounted by a factor of $(1 + r)^{-t}$, where the per-year discount rate r is set at 3.5%. This means that costs for health state $n = 1, 2, 3$ at time t are $C_{1nt} = (c_z + cc_n)(1 + r)^{-t}$ for the zidovudine monotherapy option $(d = 1)$, and $C_{2nt} = (c_z + c_l + cc_n)(1 + r)^{-t}$ for the zidovudine/lamivudine combination therapy option $(d = 2)$. Costs for state 4 (death) are zero. The health effect of interest for this decision problem is life years, so E_{dn0} is defined as 1 for health states $n = 1, 2, 3$, and zero for the death state $n = 4$. Life years accruing in the future are discounted, giving $E_{dnt} = E_{dn0}(1 + r)^{-t}$.

4. Discrepancy analysis.

4.1. Incorporating judgments about model structural error into the Markov model. We restrict ourselves to considering only structural error that relates to the Markov model itself. In many applications a Markov model is part of a larger model that may also include, for example, a decision tree element where we may also judge there to be structural error.

We believe that the transition of individuals through health states is not adequately described by a simple time-homogeneous Markov model and therefore expect there to be error in the prediction of our base case model. We wish to quantify this structural error to determine

whether we need to build a more complex model. In particular we wish to determine the expected value of improving the model.

We introduce a series of discrepancy terms, each of which represents the difference between the output of a subfunction in the built model and the true value of that output quantity. Discrepancy terms are incorporated in the model at the level of the evolution of the health state vector, replacing (3.6) with

$$(4.1) \qquad \boldsymbol{\pi}'_{dt} = \boldsymbol{\pi}'_{d,t-1}\boldsymbol{M}_d + \boldsymbol{\delta}_{dt},$$

where $\boldsymbol{\delta}_{dt}$ is a vector of discrepancy terms that quantifies the error in the state vector at time t for decision option d.

In the analysis for our case study we have found it more intuitive to think about discrepancies as applying to the transition matrix rather than to the state vector, writing $\boldsymbol{\delta}_{dt} = \boldsymbol{\pi}'_{d,t-1}\boldsymbol{\Delta}_{dt}$ and expressing judgments about the model error via $\boldsymbol{\Delta}_{dt}$, a matrix of discrepancy terms of the same dimensionality as $\boldsymbol{M}_d$. We re-express (4.1) as

$$\begin{aligned} \boldsymbol{\pi}'_{dt} &= \boldsymbol{\pi}'_{d,t-1}\boldsymbol{M}_d + \boldsymbol{\delta}_{dt} \\ &= \boldsymbol{\pi}'_{d,t-1}\boldsymbol{M}_{dt} + \boldsymbol{\pi}'_{d,t-1}\boldsymbol{\Delta}_{dt} \\ &= \boldsymbol{\pi}'_{d,t-1}(\boldsymbol{M}_d + \boldsymbol{\Delta}_{dt}) \\ (4.2) \qquad &= \boldsymbol{\pi}'_{d,t-1}\boldsymbol{M}^*_d. \end{aligned}$$

The matrix $\boldsymbol{M}^*_d$ must obey the same constraints as $\boldsymbol{M}_d$; i.e., all elements must lie within the interval $[0, 1]$, and each row must sum to one. We can ensure this if each element of $\boldsymbol{\Delta}_{dt}$, δ_{dtxy} is constrained to lie in the interval $[-p_{dxy}, 1 - p_{dxy}]$, and if each row of $\boldsymbol{\Delta}_{dt}$ sums to zero. Alternatively, we may choose to transform the elements of $\boldsymbol{M}^*_d$ onto the real line, for example, onto the log-odds scale via an elementwise logit function $g(\cdot)$, and then back-transform via $g^{-1}(\cdot)$ replacing $(\boldsymbol{M}_d + \boldsymbol{\Delta}_{dt})$ with $g^{-1}\{g(\boldsymbol{M}_d) + \boldsymbol{\Delta}_{dt}\}$. This ensures that the elements of $\boldsymbol{M}^*_d$, which we denote by p^*_{dtxy}, each lie in $[0, 1]$. We would then ensure the sum-to-one constraint for row x by setting, say, $p^*_{dtxN} = 1 - \sum_{y=1}^{N-1} p^*_{dtxy}$.

Given the transition probability matrices (3.4) and (3.5), there are potentially six such unconstrained discrepancy terms per decision option per time step, and we denote these by $\delta_{d1t}, \ldots, \delta_{d6t}$. The discrepancy matrix $\boldsymbol{\Delta}_{dt}$ is therefore

$$(4.3) \qquad \boldsymbol{\Delta}_{dt} = \begin{pmatrix} -(\delta_{d1t} + \delta_{d2t} + \delta_{d3t}) & \delta_{d1t} & \delta_{d2t} & \delta_{d3t} \\ 0 & -(\delta_{d4t} + \delta_{d5t}) & \delta_{d4t} & \delta_{d5t} \\ 0 & 0 & -\delta_{d6t} & \delta_{d6t} \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Here we judge that structural error relates only to a subset of the transitions in the model, and where we judge there to be no structural error the corresponding discrepancy term is zero.

4.2. Case study scenario 1—Time-dependent transition probabilities. In the first scenario of our case study we judge that there is an important time-dependent relationship between age and the probability of death that is not captured in the simple time-homogeneous
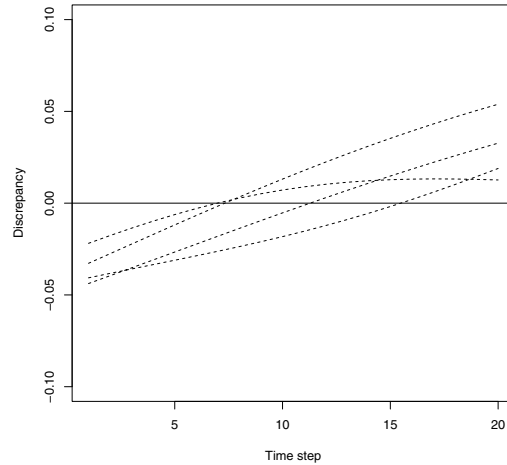
Figure 3. *Four plausible realizations of the discrepancy term δ_{16t} in scenario 1.*

model. We therefore introduce three discrepancy terms (per time step per decision), one for each transition from an alive state to the death state. Given the general expression for the discrepancy matrix in (4.3), we expect that δ_{dit} is nonzero for $i = 3, 5, 6$, and zero for $i = 1, 2, 4$. Three discrepancy terms per decision option per time step implies that there are $3 \times 2 \times 21 = 126$ discrepancy terms in total. Specifying judgments about the model discrepancy via the joint distribution of such a large number of terms clearly requires a parsimonious parametrization that reflects the dependencies between discrepancy terms.

To illustrate our approach to this specification problem we consider the discrepancy term δ_{d6t} that describes the structural error in the built model with respect to the probability of transition from AIDS to death. We judge that the probability of this transition increases monotonically over time rather than being constant in the base case model, but we are unsure as to the exact nature of the relationship between the probability of death and time. This belief implies that the uncertain discrepancy term δ_{d6t} must also increase monotonically with respect to time. We judge that at $t = 0$ the probability of death may be approximately 20% lower than the constant value (0.25) in the built model, and at $t = 20$ may be approximately 20% higher, but we have considerable uncertainty. Figure 3 represents some plausible realizations of the discrepancy δ_{d6t} as a function of time for $d = 1$.

4.3. Parametrizing the discrepancy using a Gaussian process. We wish to find a convenient and parsimonious parametrization for the joint distribution of the 126 discrepancy terms δ_{dit}, $d = 1, 2$, $i = 3, 5, 6$, and $t = 0, \ldots, 20$. We begin by noting that the reason for choosing a Markov model structure for our built model was to reflect a dynamic time-dependent process, so it seems reasonable to consider discrepancy as a function of time step, i.e., $\delta_{dit} = f_{di}(t)$. We then assume that the functions $f_{di}(t)$ follow a Gaussian process, i.e., that $\{f_{11}(0), \ldots, f_{26}(20)\}$ has a multivariate normal distribution with mean function $\mathbb{E}\{f_{di}(t)\} = \mu(d, i, t)$ and covariance function $\mathrm{Cov}\{f_{di}(t), f_{d^*i^*}(t^*)\} = \rho(d, i, t, d^*, i^*, t^*)$.

This highly flexible and parsimonious parametrization of a set of unknown functions allows us to specify not only our uncertainty about each δ_{dit}, but also the correlation structure of

discrepancies through time, the correlations between the three nonzero discrepancy terms per decision, and the correlation between the discrepancy terms for the $d = 1, 2$ decisions for each transition i.

We recognize that since we do not have discrepancies at noninteger values of t, the function $f_{di}(t)$ is discrete in this setting. Strictly speaking we are therefore defining a multivariate normal distribution for δ_{dit}, rather than a Gaussian process for $f_{di}(t)$. However, we believe that it is more natural to think about the underlying discrepancy process as being continuous in time, and hence we use the Gaussian process terminology.

4.3.1. Specifying the mean function. We specify the mean for each discrepancy, $\mathbb{E}\{f_{di}(t)\}$, as a function $\mu(d, i, t)$. For scenario 1 a linear form, $\mathbb{E}\{f_{di}(t)\} = \mu(d, i, t) = \beta_{0,di} + \beta_{1,di}t$, adequately reflects our judgments, but, depending on the decision problem, alternative choices might be a higher-order polynomial, $\mathbb{E}\{f_{di}(t)\} = \mu(d, i, t) = \beta_{0,di} + \cdots + \beta_{n,di}t^n$; a stepped polynomial, $\mathbb{E}\{f_{di}(t)\} = \mu(d, i, t) = \beta_{0,di} + \beta_{1,di}I(t > \beta_{2,di})$; or an exponential form, $\mathbb{E}\{f_{di}(t)\} = \mu(d, i, t) = \beta_{0,di} + \beta_{1,di}\exp(\beta_{2,di}t)$.

4.3.2. Specifying the covariance function. We make a number of simplifying assumptions when specifying the covariance function but note that all of these assumptions may be relaxed at the cost of specifying a greater number of hyperparameters. We assume in scenario 1 that the variance of each discrepancy δ_{dit} remains constant for all t, requiring the specification of $2 \times 3 = 6$ variances, which we denote by σ_{di}^2. We state beliefs about the within-decision, between-transition term correlation through a parameter $\phi_{ii^*} = \text{Cor}(\delta_{dit}, \delta_{di^*t})$, assuming that this is constant over time t and across decisions. We state beliefs about the between-decision correlation through a parameter $\psi_{dd^*} = \text{Cor}(\delta_{dit}, \delta_{d^*it})$, assuming that this is constant over time t and across transitions i.

Finally, we state beliefs about the correlation of the discrepancies through time by defining a correlation function $\gamma(\cdot, \cdot)$ that depends on the distance between time steps, assuming this holds for all d and i. For the purposes of scenario 1 we use the "Gaussian form"

$$(4.4) \qquad \gamma(t, t^*) = \exp\left\{-\left(\frac{t - t^*}{\omega}\right)^2\right\},$$

where ω is the correlation length. The correlation length determines the degree of correlation between discrepancy terms at any particular "distance," where distance is the number of Markov time steps between the terms. See [13] for a discussion of alternatives to this simple Gaussian form of correlation function.

The overall covariance function is therefore

$$\begin{aligned}
\text{Cov}\{f_{di}(t), f_{d^*i^*}(t^*)\} &= \rho(d, i, t, d^*, i^*, t^*) \\
&= \sigma_{di}\sigma_{d^*i^*}\psi_{dd^*}\phi_{ii^*}\rho(t, t^*) \\
&= \sigma_{di}\sigma_{d^*i^*}\psi_{dd^*}\phi_{ii^*}\exp\left\{-\left(\frac{t - t^*}{\omega}\right)^2\right\}.
\end{aligned}$$
$$(4.5)$$

4.3.3. Choosing mean and covariance function hyperparameters. Hyperparameters are chosen such that the resulting realizations of the Gaussian process represent plausible discrepancy trajectories with respect to time. This is an iterative process in which we alternate

between generating small numbers of discrepancy trajectories and adjusting ("tuning") hyperparameters until we are satisfied that the trajectories represent our judgments about the discrepancies.

We placed normal distributions on the linear mean function parameters $\beta_{0,di}$ and $\beta_{1,di}$. We began by setting the mean for $\beta_{0,di}$ to be $-1/10$ times the mean of the corresponding transition probability, and $\beta_{1,di}$ to be $-1/10$ times $\beta_{0,di}$. Values of σ_{di} were initially set equal to $\beta_{0,di}$. The correlation parameters ϕ and ψ were set at 0.8 and 0.9, respectively.

We wish to ensure that $f_{di}(t)$ is monotone with respect to t to reflect our belief that the probability of death increases with time. However, realizations of a Gaussian process tend to be "wiggly" nonmonotone functions, with the degree of "wiggliness" controlled by the ω parameter. Increasing values of ω will result in an increasingly smooth function, so by carefully choosing ω we can ensure that the realizations of the Gaussian process reflect a plausible relationship between the discrepancy and time. There are several methods for ensuring monotonicity in this context. One option would be to model the log of the increment $f_{di}(t) - f_{di}(t-1)$. Another would be to generate multiple realizations of $f_{di}(t)$, and keep only those that are monotone. Our approach is to derive the distribution of the first derivative of $f_{di}(t)$ and choose ω to ensure (with some prespecified probability) that $\partial f_{di}(t)/\partial t > 0 \ \forall t$ for a monotonically increasing function, or $\partial f_{di}(t)/\partial t < 0 \ \forall t$ for a decreasing function. Details are given in the appendix.

We generated realizations of the discrepancy trajectories over time, and if these were not plausible, we adjusted the hyperparameters and generated new realizations. We cycled through this iterative process until the discrepancy trajectories represented our judgment that at $t = 0$ the probability of death may be approximately 20% lower than the constant value (0.25) in the built model, and at $t = 20$ may be approximately 20% higher.

The hyperparameters for scenario 1 are shown in Tables 3 and 4. Ten samples from the Gaussian process for discrepancy term δ_{16t} are shown in Figure 4. Note the variation in functional form generated by the Gaussian process, reflecting our uncertainty about the relationship between probability of death and time, but with the constraint that the relationship between discrepancy and time should be monotone.

We examined similar figures for the remaining discrepancy terms to ensure that in each case the hyperparameters chosen were producing discrepancies that reflected our judgments about potential model error.

4.4. Case study scenario 2—An uncertain relationship between efficacy and time since treatment commencement. The duration of the effect of the combination therapy was a key uncertainty at the time of publication of [3], and the authors presented results for three alternative scenarios: effectiveness lasting one year, two years, and 20 years. We ask the following question: If our built model assumes that the combination therapy is effective over 20 years, but we are uncertain whether this is true, do we need to build a more complex model that incorporates an uncertain relationship between efficacy and time from commencement of treatment?

The treatment effect acts on six unconstrained terms in the transition matrix for the combination therapy (3.5) but does not act on the transition matrix for the monotherapy, therefore resulting in six nonzero discrepancies per time step, $\delta_{21t}, \ldots, \delta_{26t}$. This specification

<div align="center">

Table 3

Hyperparameters to specify GP mean function.

</div>

| Parameter	Scenario 1	Scenario 2	Scenario 3	Transition
Intercept $(\beta_{0,di})$	Mean (SD) $\times 10^{-3}$	Mean (SD) $\times 10^{-3}$	Mean (SD) $\times 10^{-3}$	State x to y
$\beta_{0,11}$	0 (0)	0 (0)	0 (0)	1 to 2
$\beta_{0,12}$	0 (0)	0 (0)	0 (0)	1 to 3
$\beta_{0,13}$	-1.0 (0.20)	0 (0)	0 (0)	1 to 4
$\beta_{0,14}$	0 (0)	0 (0)	0 (0)	2 to 3
$\beta_{0,15}$	-1.2 (0.24)	0 (0)	0 (0)	2 to 4
$\beta_{0,16}$	-25.0 (5.00)	0 (0)	0 (0)	3 to 4
$\beta_{0,21}$	0 (0)	0 (0)	0 (0)	1 to 2
$\beta_{0,22}$	0 (0)	0 (0)	0 (0)	1 to 3
$\beta_{0,23}$	-0.5 (0.10)	0 (0)	0 (0)	1 to 4
$\beta_{0,24}$	0 (0)	0 (0)	0 (0)	2 to 3
$\beta_{0,25}$	-0.6 (0.12)	0 (0)	0 (0)	2 to 4
$\beta_{0,26}$	-12.7 (2.54)	0 (0)	0 (0)	3 to 4
Slope $(\beta_{1,di})$	Mean (SD) $\times 10^{-4}$	Mean (SD) $\times 10^{-4}$	Mean (SD) $\times 10^{-4}$	State x to y
$\beta_{1,11}$	0 (0)	0 (0)	0 (0)	1 to 2
$\beta_{1,12}$	0 (0)	0 (0)	0 (0)	1 to 3
$\beta_{1,13}$	1.0 (0.20)	0 (0)	0 (0)	1 to 4
$\beta_{1,14}$	0 (0)	0 (0)	0 (0)	2 to 3
$\beta_{1,15}$	1.2 (0.24)	0 (0)	0 (0)	2 to 4
$\beta_{1,16}$	25.0 (5.00)	0 (0)	0 (0)	3 to 4
$\beta_{1,21}$	0 (0)	24.8 (13.78)	0 (0)	1 to 2
$\beta_{1,22}$	0 (0)	8.2 (4.57)	0 (0)	1 to 3
$\beta_{1,23}$	0.5 (0.10)	1.2 (0.68)	0 (0)	1 to 4
$\beta_{1,24}$	0 (0)	50.0 (27.76)	0 (0)	2 to 3
$\beta_{1,25}$	0.6 (0.12)	1.5 (0.82)	0 (0)	2 to 4
$\beta_{1,26}$	12.7 (2.54)	30.7 (17.0)	0 (0)	3 to 4

of discrepancy is equivalent to incorporating a time-varying treatment effect parameter (the treatment effect parameter α is constant in the base case model), but with the additional flexibility that allows the treatment effect to vary across the different transitions in the model (e.g., HIV to AIDS versus HIV to death).

We believe that efficacy falls over time, and therefore that the discrepancy between our built model and reality increases over time. We again chose a linear mean function $\mathbb{E}(\delta_{2it}) = \beta_{0,i} + \beta_{1,i}t$ with uncertain slope. The intercept parameter $\beta_{0,i}$ is zero in this case to reflect our judgment that during time step 1 the treatment effect parameter α correctly determines the effectiveness of the combination therapy. We placed normal distributions on the six $\beta_{1,i}$ parameters with hyperparameters shown in Table 3.

Next, we specify the covariance function. Our uncertainty about the six discrepancies $\delta_{21t}, \ldots, \delta_{26t}$ is controlled through variance terms $\sigma_{21}^2, \ldots, \sigma_{26}^2$, assumed to hold for all t. We specify our judgment about the dependency between the discrepancy terms for the six transitions through a single correlation parameter $\phi_{ii^*} = \phi \; \forall i \neq i^*$, which we assume constant for all t. Since there is no discrepancy for the monotherapy option $d = 1$ in this scenario, we do

Variance parameters (σ_{di})	Scenario 1 $(\times 10^{-3})$	Scenario 2 $(\times 10^{-3})$	Scenario 3 $(\times 10^{-3})$	Transition state x to y
σ_{11}	0	0	28.9	1 to 2
σ_{12}	0	0	9.6	1 to 3
σ_{13}	1.0	0	1.4	1 to 4
σ_{14}	0	0	58.1	2 to 3
σ_{15}	1.2	0	1.7	2 to 4
σ_{16}	25.0	0	35.7	3 to 4
σ_{21}	0	5.1	14.7	1 to 2
σ_{22}	0	1.7	4.9	1 to 3
σ_{23}	0.51	0.25	0.7	1 to 4
σ_{24}	0	10.4	29.6	2 to 3
σ_{25}	0.61	0.31	0.9	2 to 4
σ_{26}	12.7	6.4	18.1	3 to 4

| Correlation parameters	Scenario 1	Scenario 2	Scenario 3	Description
ϕ	0.8	0.9	0	between discrepancy term correlation
ψ	0.9	0	0	between-decision correlation
ω	32	7	7	correlation length parameter
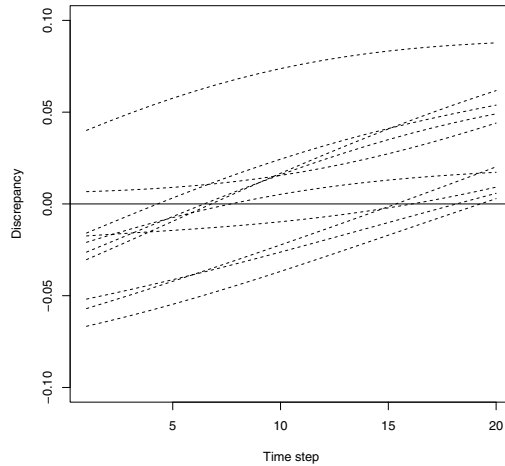
Figure 4. *Ten samples from the distribution on discrepancy term δ_{16t} in scenario* 1.

not need to specify between-decision correlations (i.e., there is no ψ_{dd*} correlation parameter). Finally, we specify a correlation structure for the discrepancies as they evolve through time via a Gaussian form correlation function (4.4) with parameter ω, ensuring via (A.6) that discrepancy as a function of time is monotone with probability $\alpha = 0.95$. Values for all covariance function parameters are shown in Table 4.

Ten samples from the Gaussian process for discrepancy term δ_{21t} are shown in Figure 5.
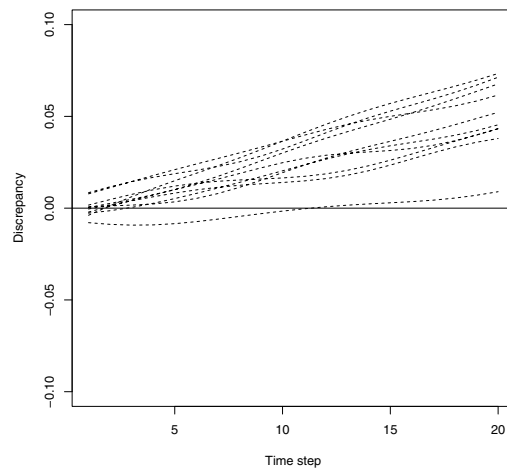
Figure 5. *Ten samples from the distribution on discrepancy term δ_{21t} in scenario* 2.

Note the variation in functional form generated by the Gaussian process, reflecting our uncertainty about the relationship between efficacy and time. Again, we examined similar figures for the remaining discrepancy terms to ensure that in each case the hyperparameters chosen were producing discrepancies that reflected our judgments about potential model error.

4.5. Case study scenario 3—Relaxation of the memoryless property. In scenario 3 we judge that the probability of transition to state y at time step $t + 1$ is dependent not only on the state x at time t but on the states occupied at time steps $\leq t - 1$. We therefore want to relax the Markov assumption and consider more complex time dependencies that would necessitate a more flexible modelling framework (for example, using a discrete event or agent based approach). In order to judge whether this is necessary we add relatively unstructured discrepancy to allow for a wide range of possible deviations from the simple memoryless Markov process. Hyperparameters are shown in Tables 3 and 4. Ten samples from the Gaussian process for discrepancy term δ_{21t} are shown in Figure 6. Again, we examined similar figures for the remaining discrepancy terms to ensure plausibility.

5. Results.

5.1. Base case model with no discrepancy. We implemented the model in R [15]. We sampled from the model input parameters and ran the model 10,000 times. Mean costs were £51,394 and £96,807 for the monotherapy and combination therapy options, respectively, with corresponding mean health effects of 6.60 and 10.47 life years. With a value of $\lambda = \pounds12,000$ per life year[1] this results in predicted net benefits of £27,829.51 and £28,825.06, and therefore is an indication that the combination therapy option should be preferred. The incremental analysis (which calculates the difference in costs per unit difference in health effects) suggests that by adopting combination therapy rather than monotherapy each additional life year would be

---

[1]We assumed for the purposes of this case study a willingness to pay of $\lambda = \pounds12,000$ per life year to ensure that we were in the region of decision uncertainty. This is lower than the value that would be used for decisions in many Western health economies.
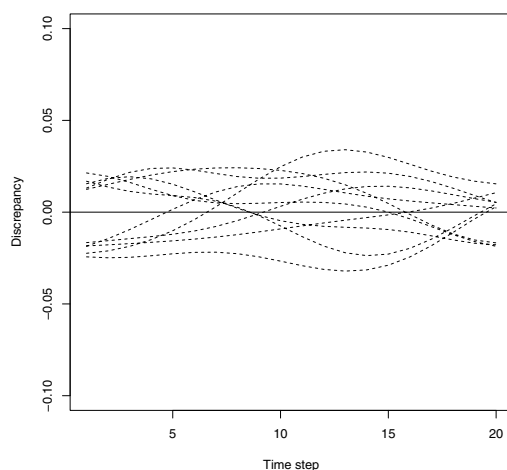
Figure 6. *Ten samples from the distribution on discrepancy term δ_{21t} in scenario* 3.

Table 5
Partial EVPI results.

Parameter	Partial EVPI (EVPI index[†])			
	Base case	Scenario 1	Scenario 2	Scenario 3
Transition probabilities	£0.00 (0.0%)	£0.00 (0.0%)	£0.00 (0.0%)	£1.17 (0.1%)
Relative risk	£169.91 (46.5%)	£227.37 (50.7%)	£64.63 (19.4%)	£164.55 (17.2%)
Costs	£194.41 (53.2%)	£215.56 (48.0%)	£65.17 (19.55%)	£167.53 (17.5%)
Discrepancy terms	–	£16.87 (3.8%)	£110.21 (33.1%)	£699.06 (73.0%)
Overall EVPI	£365.42	£448.33	£333.43	£957.28

† The partial EVPI as a proportion of the overall EVPI.

gained at a cost of £11,742. However, there is decision uncertainty, which is reflected in the overall value of learning the input parameters of £365.42. The results in Table 5 suggest that decision uncertainty is being driven by uncertainty in the treatment effect parameter with EVPI(relative risk) = £169.91 (EVPI index, 46.5%) and uncertainty in the cost parameters with EVPI(costs) = £194.41 (53.2%).

5.2. Scenario 1. After the addition of discrepancy to reflect the judgments about model error due to the time homogeneity assumption, mean net benefits were £27,997 and £28,844, indicating that the combination therapy option is still optimal. Incremental analysis results in a cost per life year gained of £11,776. Value of information analysis suggests that the decision uncertainty is still dominated by the uncertainty in the inputs with EVPI(relative risk) = £227.37 (50.7%) and EVPI(costs) = £215.56 (48.0%). There is little value in learning $\boldsymbol{\delta}$ with EVPI($\boldsymbol{\delta}$) = £16.87 (3.8%), indicating that building a more complex model is not advisable at a willingness to pay for one life year of $\lambda = £12,000$ (Table 5). It appears that uncertainty regarding the model error that results from the time homogeneity assumption is not a significant driver of decision uncertainty. This would suggest that our simple built model is "good

enough" for the decision in this scenario.

5.3. Scenario 2. After the addition of discrepancy terms to reflect the judgments about model error due to the constant treatment efficacy assumption, mean net benefits were £27,767 and £26,466, indicating that the monotherapy option is now optimal. Incremental analysis results in a cost per life year gained of £12,407. Value of information analysis suggests that although there is still some value in learning the treatment effect and cost parameters, it is the discrepancy terms that are now most important in driving decision uncertainty (Table 5). In this scenario there is value in improving the model such that it better reflects our judgments about the decision problem, as well as value in reducing parameter uncertainty.

5.4. Scenario 3. After the addition of discrepancy terms to reflect the judgments about model error due to the Markovian assumption of memorylessness, mean net benefits were £28,071 and £28,982, indicating that the combination therapy option is optimal. Incremental analysis results in a cost per life year gained of £11,762. Value of information analysis suggests that the decision is again sensitive to the discrepancy terms, and that building a more complex model to better represent non-Markovian transitions between health states may be worthwhile (Table 5).

6. Discussion. We have demonstrated a method for incorporating within a model judgments about the structural error that results from building an "incorrect" model. The method allows us to determine the expected value of building a better model, given the uncertainty in the input parameters. This approach will be most valuable in cases where the decision problem is complex, but due to difficulties in obtaining input parameter estimates or lack of time or resources we have built a simple model. We feel that this may be of particular relevance in the emerging field of economic evaluation of public health interventions where decision problems generally have many complex elements, but models are often relatively simple (for good examples see descriptions of the models that have been used by the National Institute for Health and Care Excellence to support public health intervention resource allocation decisions in England (http://www.nice.org.uk/Guidance/PHG/Published)).

In health economic applications the objective is typically to choose between discrete treatment options (i.e., to optimize over a single factor), and in most cases the utility function for the predicted response is well defined (or even mandated by an external body). In the engineering and physical sciences a more common objective is to optimize a set of parameters that are continuous variables. The form of the utility function and sensitivity analysis proposed in (2.5) can be generalized easily to accommodate this (see [17] for a review of sensitivity analysis measures).

The most important potential practical limitation of the method lies in our ability to meaningfully specify the joint distribution on the inputs and discrepancies, $p(\boldsymbol{X}, \boldsymbol{\delta})$. In our case study we represented our beliefs about $\boldsymbol{\delta}$ fairly crudely and made the assumption that inputs and discrepancies were independent. However, even if judgments are crude, as long as we are "generous" with our specification of uncertainty, the expected value of learning $\boldsymbol{\delta}$ will provide an upper bound on the value of better modelling. If EVPI($\boldsymbol{\delta}$) is small compared with the value of learning the inputs, even with the generous estimate of uncertainty about the structural error, we can be reassured that the current model is "good enough." In contrast,

if EVPI($\boldsymbol{\delta}$) dominates EVPI($\boldsymbol{X}$), then we conclude that it will be worthwhile to rebuild the model so that it better reflects our beliefs about the relationships between the inputs and the target quantities we wish to predict.

We believe that the internal discrepancy analysis method offers a promising approach to managing computer model structural uncertainty. By avoiding the requirement to make direct judgments about the external model discrepancy, we can make judgments about structural error in the absence of observations on the model output. However, in applications in which there are such observations, we expect that it should be possible to extend our approach to specify a joint probability model over inputs, discrepancies, and observations on the output. This is an area for future exploration.

Appendix. Ensuring that $f_{di}(t)$ is monotonic with respect to t. Monotonicity with respect to t implies that, for a once differentiable function $f_{di}(t)$, $\partial f_{di}(t)/\partial t > 0 \ \forall t$ or $\partial f_{di}(t)/\partial t < 0 \ \forall t$. Informally then, we can ensure monotonicity by choosing hyperparameters for the mean and covariance functions such that this holds with some probability θ [12]. We use the following property of a Gaussian process, described in [14]. If $f(x) \sim GP\{\mu(x), \rho(x, x^*)\}$ is an n times differentiable Gaussian process with n times differentiable mean and covariance functions, then $\partial^n f(x)/\partial x^n$ is also a Gaussian process with mean function

$$(A.1) \qquad \mathbb{E}\left\{\frac{\partial^n}{\partial x^n} f(x)\right\} = \frac{\partial^n}{\partial x^n}\mu(x)$$

and covariance function

$$(A.2) \qquad \text{Cov}\left\{\left.\frac{\partial^n}{\partial x^n} f(x)\right|_{x=x}, \left.\frac{\partial^n}{\partial x^n} f(x)\right|_{x=x^*}\right\} = \frac{\partial^{2n}}{\partial x^n \partial x^{*n}}\rho(x, x^*).$$

This implies that $\partial f_{di}(t)/\partial t$ is the Gaussian process

$$(A.3) \qquad \frac{\partial}{\partial t} f_{di}(t) \sim GP\left\{\frac{\partial}{\partial t}\mu(d, i, t), \frac{\partial^2}{\partial t \partial t^*}\rho(d, i, t, d, i, t^*)\right\},$$

and we can ensure monotonicity of $f_{di}(t)$ with some prespecified probability θ by choosing parameters of $m(\cdot)$ and $\rho(\cdot, \cdot)$ such that

$$(A.4) \qquad \left|\frac{\partial}{\partial t}\mu(d, i, t)\right| - \Phi^{-1}(\theta)\sqrt{\frac{\partial^2}{\partial t \partial t^*}\rho(d, i, t, d, i, t^*)} > 0,$$

where $\Phi^{-1}(\cdot)$ is the inverse normal cumulative distribution function. Given a linear mean function,

$$\mu(d, i, t) = \beta_{0,di} + \beta_{1,di}t,$$

and a Gaussian form for the correlation function with respect to time,

$$\rho(d, i, t, d, i, t^*) = \sigma_{di}^2 \exp\left\{-\left(\frac{t - t^*}{\omega}\right)^2\right\},$$

expression (A.4) becomes

$$(\text{A.5}) \qquad \left| \beta_{1,di} \right| - \Phi^{-1}(\theta)\sqrt{\frac{2\sigma_{di}^2}{\omega^2}} > 0,$$

which by solving for ω gives

$$(\text{A.6}) \qquad \omega > \frac{\Phi^{-1}(\theta)\sqrt{2}\sigma_{di}}{\beta_{1,di}}.$$

We can therefore, given $\beta_{1,di}$ and σ_{di}^2, ensure with some probability θ that $f_{di}(t)$ is monotone through a choice of correlation length parameter ω that obeys (A.6). We set $\theta = 0.95$.

REFERENCES

[1] J. M. BERNARDO AND A. F. M. SMITH, *Bayesian Theory*, John Wiley & Sons, Chichester, UK, 1994.

[2] A. BRIGGS, K. CLAXTON, AND M. SCULPHER, *Decision Modelling for Health Economic Evaluation*, Oxford University Press, Oxford, UK, 2006.

[3] J. V. CHANCELLOR, A. M. HILL, C. A. SABIN, K. N. SIMPSON, AND M. YOULE, *Modelling the cost effectiveness of lamivudine/zidovudine combination therapy in HIV infection*, PharmacoEconomics, 12 (1997), pp. 54–66.

[4] P. S. CRAIG, M. GOLDSTEIN, J. C. ROUGIER, AND A. H. SEHEULT, *Bayesian forecasting for complex systems using computer simulators*, J. Amer. Statist. Assoc., 96 (2001), pp. 717–729.

[5] D. DRAPER, *Assessment and propagation of model uncertainty*, J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 45–97.

[6] M. F. DRUMMOND, M. J. SCULPHER, G. W. TORRANCE, B. O'BRIEN, AND G. L. STODDART, *Methods for the Economic Evaluation of Health Care Programmes*, 3rd ed., Oxford University Press, Oxford, UK, 2005.

[7] M. GOLDSTEIN AND J. ROUGIER, *Reified Bayesian modelling and inference for physical systems*, J. Statist. Plann. Inference, 139 (2009), pp. 1221–1239.

[8] D. HIGDON, M. KENNEDY, J. C. CAVENDISH, J. A. CAFEO, AND R. D. RYNE, *Combining field data and computer simulations for calibration and prediction*, SIAM J. Sci. Comput., 26 (2004), pp. 448–466.

[9] J. B. KADANE AND N. A. LAZAR, *Methods and criteria for model selection*, J. Amer. Statist. Assoc., 99 (2004), pp. 279–290.

[10] R. E. KASS AND A. E. RAFTERY, *Bayes factors*, J. Amer. Statist. Assoc., 90 (1995), pp. 773–795.

[11] M. C. KENNEDY AND A. O'HAGAN, *Bayesian calibration of computer models*, J. Roy. Statist. Soc. Ser. B, 63 (2001), pp. 425–464.

[12] R. MONTES DIEZ AND J. E. OAKLEY, *Gaussian Processes Priors for Monotone Functions*, unpublished manuscript, 2010.

[13] R. M. NEAL, *Regression and classification using Gaussian process priors*, in Bayesian Statistics, Vol. 6, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., Oxford University Press, New York, 1999, pp. 475–501.

[14] A. O'HAGAN, *Some Bayesian numerical analysis*, in Bayesian Statistics, Vol. 4, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., Oxford University Press, New York, 1992, pp. 345–363.

[15] R DEVELOPMENT CORE TEAM, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011.

[16] H. RAIFFA, *Decision Analysis. Introductory Lectures on Choices under Uncertainty*, Addison-Wesley, Reading, MA, 1968.

[17] A. SALTELLI, M. RATTO, T. ANDRES, F. CAMPOLONGO, J. CARIBONI, D. GATELLI, M. SAISANA, AND S. TARANTOLA, *Global Sensitivity Analysis: The Primer*, John Wiley & Sons, Chichester, UK, 2008.

[18] M. STRONG, J. E. OAKLEY, AND J. CHILCOTT, *Managing structural uncertainty in health economic decision models: A discrepancy approach*, J. Roy. Statist. Soc. Ser. C, 61 (2012), pp. 25–45.