Published paper

Strong, M., Oakley, J., Chilcott, J. (2012) *Managing structural uncertainty in health economic decision models: a discrepancy approach*, Journal of the Royal Statistical Society, Series C (Applied Statistics), 61 (1), pp. 25-45
http://dx.doi.org/10.1111/j.1467-9876.2011.01014.x

Managing structural uncertainty in health economic decision models: a discrepancy approach

Mark Strong[1,§], Jeremy E. Oakley[2] and Jim Chilcott[1]

1. School of Health and Related Research (ScHARR), University of Sheffield, UK.

2. School of Mathematics and Statistics, University of Sheffield, UK.

[§] Corresponding author m.strong@sheffield.ac.uk

September 2011

The published version of this paper is: Strong M, Oakley JE, Chilcott, J. Managing structural uncertainty in health economic decision models: a discrepancy approach. Journal of the Royal Statistical Society, Series C. 2012;61(1):25-45. Available at
http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9876.2011.01014.x/abstract.

Abstract

Healthcare resource allocation decisions are commonly informed by computer model predictions of population mean costs and health effects. It is common to quantify the uncertainty in the prediction due to uncertain model inputs, but methods for quantifying uncertainty due to inadequacies in model structure are less well developed. We introduce an example of a model that aims to predict the costs and health effects of a physical activity promoting intervention. Our goal is to develop a framework in which we can manage our uncertainty about the costs and health effects due to deficiencies in the model structure. We describe the concept of 'model discrepancy': the difference between the model evaluated at its true inputs, and the true costs and health effects. We then propose a method for quantifying discrepancy based on decomposing the cost-effectiveness model into a series of sub-functions, and considering potential error at each sub-function. We use a variance based sensitivity analysis to locate important sources of discrepancy within the model in order to guide model refinement. The resulting improved model is judged to contain less structural error, and the distribution on the model output better reflects our true uncertainty about the costs and effects of the intervention.

KEYWORDS: computer model, elicitation, health economics, model uncertainty, sensitivity analysis, uncertainty analysis

1 Introduction

Mathematical "cost-effectiveness" models are routinely used to aid healthcare resource allocation decisions. Such models estimate the population mean costs and health effects of a range of decisions, and will be most helpful when their results are unbiased and uncertainty about their estimated costs and consequences is properly specified. Two sources of uncertainty in model predictions are uncertainty about the model *input* values and uncertainty about model *structure*.

These models are typically 'law-driven' (based on our knowledge of the system) rather than 'data-driven' (fitted to data), following the distinction given in Saltelli et al. (2008). Indeed, such models are built because of a lack of data on long term costs and health consequences. The law-driven nature of the cost-effectiveness model has important implications for our choice of technique for managing structural uncertainty, as we discuss later.

To quantify input uncertainty, one can specify a probability distribution for the true values of the inputs, and propagate this distribution through the model, typically using Monte Carlo sampling. In health economic modelling, this is known as probabilistic sensitivity analysis (PSA) (Claxton et al., 2005). The danger with reporting uncertainty based only on a PSA is that this may be interpreted as quantifying uncertainty about the costs and health effects of the various decision options. However, PSA only quantifies uncertainty in the *model output* due to uncertainty in model inputs. To properly represent uncertainty about the costs and health effects we must also consider uncertainty in the model structure. However, quantifying uncertainty in model structure is hard since it requires judgements about a model's ability to faithfully represent a complex real life decision problem.

Model averaging methods can be used to assess structural uncertainty if a complete set of plausible competing models can be built and weighted according to some measure of model adequacy. The weighting may be based, for example, on the posterior probability that the model is 'correct', or the predictive power of the model in a cross-validation framework. See Kadane and Lazar (2004) for a general discussion on this topic and Jackson et al. (2009, 2010) for more focussed discussions with respect to health economic decision model uncertainty. Model averaging does however have limitations. If model weights are dependent on observed data then we must be able to write down a likelihood function linking the model output to the data. This will be difficult unless we have observations on

the output of the model itself, which in the health economic context we almost never have. If we have observations on a surrogate endpoint (say, drug efficacy at two years in the context of wishing to predict efficacy at ten years) then we can construct weights that relate to certain structural choices within the model, but crucially the data will not guide our choice of the *whole* model structure. Hence there may be elements within each model that lead to different predictions of the output, but are untested in the model averaging framework.

The problem of model structure uncertainty has also been addressed in the computer models literature, but from a different perspective. Rather than focusing on generating weights for models within some set, methods are directed towards making inferences about model "discrepancy": the difference between the model run at its 'best' or 'true' input, and the true value of the output quantity (Kennedy and O'Hagan, 2001). Given a model, written as a function f, with (uncertain) inputs $\boldsymbol{X}$, the key expression is equation (1), which links the model output $Y = f(\boldsymbol{X})$ to the true, but unknown value of the target quantity we wish to predict, Z:

$$Z = f(\boldsymbol{X}) + \delta_z, \tag{1}$$

The discrepancy term, δ_z, quantifies the *structural error*: the difference between the output of the model evaluated at its true inputs and the true target quantity. We are explicitly recognising in equation (1) that our model may be deficient, but note that when we speak about model deficiency we are not concerned with mistakes, 'slips', 'lapses' or other errors of implementation (for a discussion on this topic see Chilcott et al., 2010b). Rather, we are concerned with deficiencies arising as a result of the gap between our model of reality, and reality itself. Obtaining a joint distribution that reflects our beliefs about inputs and discrepancies, $p(\boldsymbol{X}, \delta_z)$, allows us then to fully quantify our uncertainty in the target quantity due to both uncertain inputs and uncertain structure. This approach has the important advantage that only a single model need be built, though methods for making inferences about discrepancy in the context of multiple models have also been explored (Goldstein and Rougier, 2009).

In our paper we explore the feasibility of the discrepancy method in assessing structural uncertainty in a cost effectiveness model for a physical activity promoting intervention. In section 2 we describe our 'base case' model and report results without any assessment of structural uncertainty. In section 3 we describe our proposed method for quantifying the model discrepancy δ_z. We describe the application of the method to our model in section 4 and present results in section

5. We discuss implications and potential further developments in the final section.

2 Case study: a physical activity intervention cost effectiveness model

We based our case study on a cost-effectiveness model that was developed to support the National Institute for Health and Clinical Excellence (NICE) physical activity guidance (NICE, 2006). NICE is the organisation in the UK that makes recommendations to the National Health Service and other care providers on the clinical and cost effectiveness of interventions for promoting health and treating disease. The majority of NICE's recommendations and guidance products are informed by mathematical model predictions.

Our simplified version of the NICE model aims to predict the incremental net benefit of two competing decision options: exercise on prescription (e.g. from a general medical practitioner) to promote physical activity (the 'intervention'), and a 'do nothing' scenario ('no intervention'). Incremental net benefit, measured in monetary units, is defined as

$$Z = \lambda(E_2 - E_1) - (C_2 - C_1) = \lambda\Delta E - \Delta C, \tag{2}$$

where E_d and C_d are respectively the population mean health effects and costs associated with decisions $d = 1, 2$, and λ is the value in monetary units that the decision maker places on one unit of health effect. We assume that the intervention impacts on health by reducing the risks of three diseases: coronary heart disease (CHD), stroke and diabetes. The health effects included in the model are those that relate to these three diseases, and we count costs that accrue as a result of the treatment of the three diseases, as well as those that relate to the intervention itself.

2.1 Description of 'base case' model - no assessment of structural uncertainty

Our model is a simple static cohort model which can be viewed as a decision tree (figure 1). The left-most node represents the two decision options, $d = 1$, no intervention, and $d = 2$, the exercise prescription intervention. The first chance node represents the probability of new exercise under each decision option, with

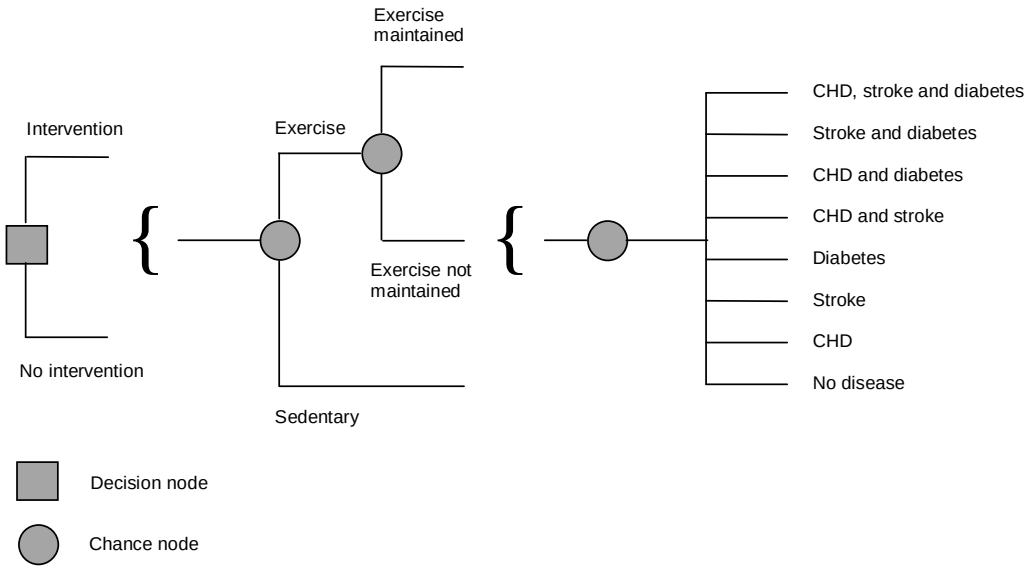

Figure 1: The model expressed as a decision tree

the second node representing the probability of maintenance of exercise conditional on new exercise. The third node represents the probability of eight mutually exclusive health states conditional on each of the three outcomes from the first two nodes: exercise that is maintained, exercise that is not maintained, and no exercise (sedentary lifestyle).

The structure of the model represents our beliefs about the causal links between the intervention and exercise, and exercise and health outcomes. There are no data available that relate to the model outputs; we have not observed costs and health outcomes for control and treatment groups on the exercise intervention. However, separate data sources are available regarding the effectiveness of the intervention in promoting exercise, and the risks of the various disease outcomes for active versus sedentary patients, and the availability of such data has guided the choice of model structure.

In our model each comorbid health state (e.g. the state of CHD *and* stroke) is treated as having a single onset point in time. Individuals do not progress, say, from the disease free state, to CHD and then to CHD plus stroke as they might do in reality. This is clearly unrealistic and is a consequence of the choice to use a very simple decision tree structure. Modelling sequential events is possible using a decision tree structure, but the number of terminal tree branches quickly becomes very large in all but the simplest of problems (Sonnenberg and Beck, 1993). A Markov or discrete event model structure would be more suited to addressing our decision problem (see Karnon (2003) for a comparison of these methods), but

we have chosen to retain the important features of the structure of the model published by NICE, upon which our case study is based (NICE, 2006).

We denote the set of eight health states, *disease free, CHD alone, stroke alone, diabetes alone, CHD and stroke, CHD and diabetes, stroke and diabetes, CHD and stroke and diabetes* as $\mathcal{H} = \{h_j, j = 1, \ldots, 8\}$, where j indexes the set in the order given above. Each of the eight health states $h_j \in \mathcal{H}$, under each decision option d, has a cost c_{dj} (measured in £), a health effect (measured in Quality Adjusted Life Years) q_{dj}, and a probability of occurrence p_{dj} (as approximated by the relative frequency with which this health state occurs within a large cohort). Total costs and total health effects for decision d are obtained by summing over health states, i.e. $C_d = \sum_{j=1}^{8} c_{dj}p_{dj}$ and $Q_d = \sum_{j=1}^{8} q_{dj}p_{dj}$. Given these, the model predicted incremental net benefit, Y is

$$Y = \lambda(Q_2 - Q_1) - (C_2 - C_1) = \lambda\Delta Q - \Delta C. \tag{3}$$

The costs c_{dj}, health effects q_{dj}, and health state probabilities p_{dj} are not themselves input parameters in the model, but instead are functions of input parameters. There are 24 uncertain and three fixed input parameters that relate to the costs, quality of life and epidemiology of CHD, stroke and diabetes, and the effectiveness of the intervention in increasing physical activity. These inputs are denoted $\boldsymbol{X} = X_1, \ldots, X_{27}$, and uncertainty is represented via the joint distribution $p(\boldsymbol{X})$. The input quantities and their distributions are described in tables 2 and 3 in appendix A.

Finally, we denote the deterministic function that links the model inputs to the model output as f, i.e. $Y = f(\boldsymbol{X})$, and call this the *base case model*.

2.2 Base case model results

The model function (which we describe in detail in section 4) was implemented in R (R Development Core Team, 2010). We sampled the input space and ran the model 100,000 times. The mean of the model output, Y, at λ=£20,000/QALY was £247 and the 95% credible interval was (-£315, £1002). The probability that the intervention is cost-effective, $P(\text{INB} > 0)$, at λ =£20,000 was 0.77. Results for the base case model are shown graphically in figure 2 (note that figure 2 also includes the results for the 'with discrepancy' and 'after remodelling' analyses that are reported in section 5.1).

Figure 2a shows the cost-effectiveness plane (with 100 Monte Carlo samples). The sloped line shows the willingness to pay threshold of £20,000 per QALY. To

aid clarity figure 2b is a contour plot representation of the cost effectiveness plane, showing the 95^{th} percentile of an empirical kernal density estimate of the joint distribution of costs and effects. Figure 2c shows the cost-effectiveness acceptability curve (i.e. a plot of $P(\text{INB} > 0)$ against λ) for values of λ from £0/QALY to £40,000/QALY. Finally, figure 2d shows the kernel density estimate for Y, the base case model estimate of the incremental net benefit at λ =£20,000.

A mean incremental net benefit of £247 at λ=£20,000/QALY implies that, on average, the intervention will accrue costs and health effects that have a positive net value of £247 per person treated. The probabilistic sensitivity analysis implies that, at λ=£20,000/QALY, a choice to recommend the intervention would have a probability of 0.77 of being better than the choice not to recommend.

3 Managing uncertainty due to structure: a discrepancy approach

For the decision maker to base their decision on the model output, the model must have credibility. The model must be judged good enough to support the decision being made. The primary goal of our analysis is therefore to provide a means for quantifying judgements about structural error and specifically to determine the relative importance of structural compared to input uncertainty in addressing the decision problem. If uncertainty about structural error is large then we may wish to review the model structure. Conversely, if we can demonstrate that the uncertainty about structural error is small in comparison to that due to input uncertainty, then we have a stronger claim to have built a credible model.

In building the base case model we made a series of assumptions, for example we assumed that occurrences of CHD, stroke and diabetes are independent at the level of the individual and therefore that disease risks act multiplicatively. Such assumptions drive the structural choices that we make when formulating a model, and incorrect assumptions will lead to structural error. We must therefore focus our attention on the assumptions within a model if we are to assess its adequacy and properly quantify our uncertainty about the target quantity.

In the model averaging framework new models would be built to incorporate the set of alternative assumptions believed plausible (with new models possibly being just minor variants of the existing model). The models would then be weighted according to some measure of adequacy in relation to data, D. Given

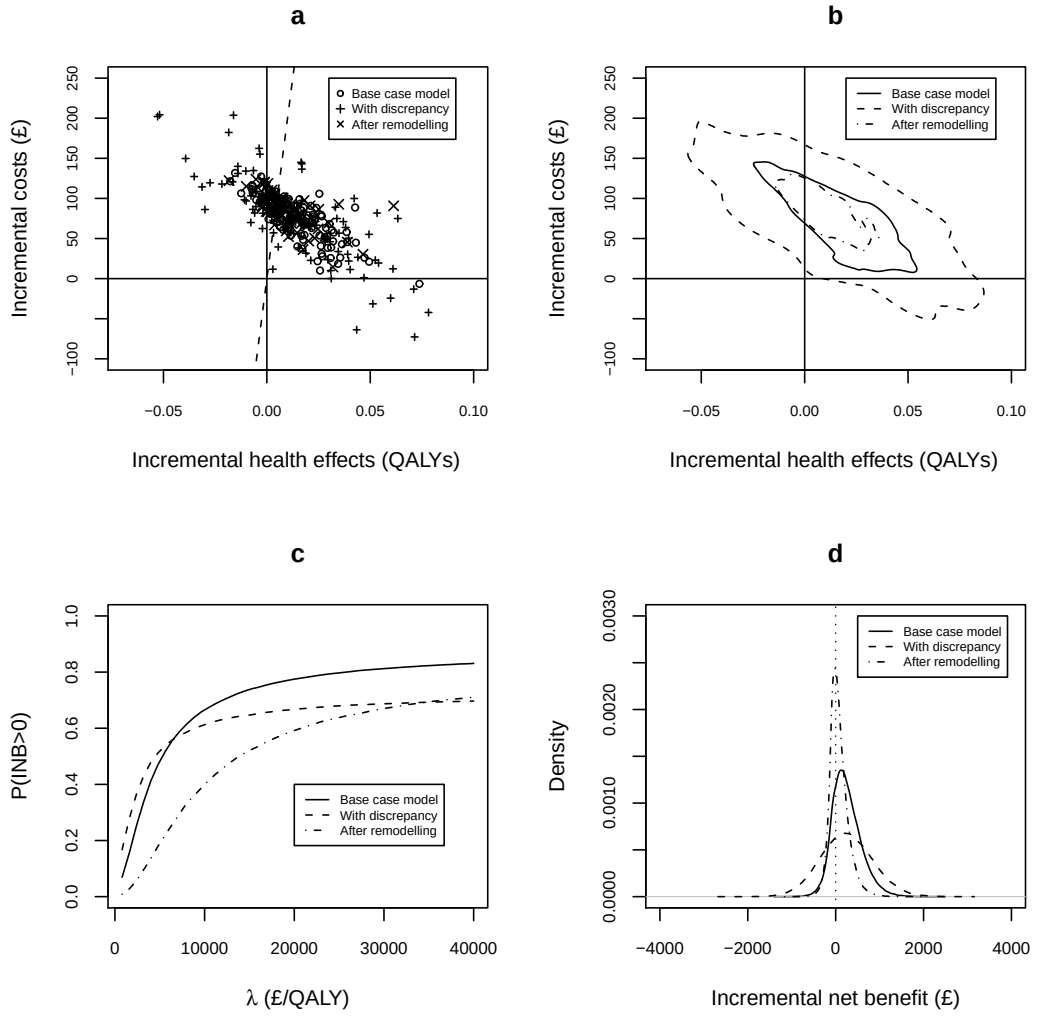

Figure 2: Model output shown as (a) cost-effectiveness plane (b) cost-effectiveness plane contour plot (c) cost-effectiveness acceptability curve (d) incremental net benefit empirical density. Results are shown for the base case model (section 2.2), 'with discrepancy' analysis (section 5.1) and 'after remodelling' analysis (section 5.6).

a set of models $\{M_i, i \in \mathcal{I}\}$ and adequacy measure $\omega(\cdot)$, the distribution of the target incremental net benefit is given by

$$p(\boldsymbol{Z}|D) = \sum_{i \in \mathcal{I}} p(\boldsymbol{Z}|M_i, D)\omega(M_i|D). \tag{4}$$

If we believe that one of the models in our set is true (i.e. that $\{M_i, i \in I\}$ is "$\mathcal{M}$-closed" in the terminology of Bernardo and Smith, 1994), and can specify prior model probabilities $p(M_i)$, then the models can be weighted by their posterior probabilities given the data,

$$p(M_i|D) = \frac{p(D|M_i)p(M_i)}{\sum_{i \in I} p(D|M_i)p(M_i)}, \tag{5}$$

For the $\mathcal{M}$-closed case this is a consistent estimation procedure, in the sense that as more data are collected the posterior probability of the true model will converge to 1. However, if we believe that none of the models is correct (i.e. we have an "$\mathcal{M}$-open" set) then this approach is no longer consistent. In the $\mathcal{M}$-open case Jackson et al. (2010) propose instead that weights are based on the predictive probability of M_i given a replicate data set.

A more fundamental problem in the context of health economic decision modelling is the usual absence of data against which to measure the adequacy of the model in its entirety. We do not measure overall costs and health effects over extended time periods under competing decision options. In the absence of observations on the model output $\boldsymbol{Z}$, weights could be based on the judgement of the modeller and/or decision maker, though making probability statements about models, which are by definition abstract non-observables is likely to be very difficult.

We therefore propose a different approach based on specifying a distribution for the model discrepancy, δ_z, as defined in equation (1). In contrast to the model averaging approach we do not attempt to make assessments about the adequacy of the model structure in relation to alternative structures; we instead assess how large an error might be due to the structure of the model at hand.

3.1 Discrepancy between model output and reality

In many applications in the physical sciences the target quantity predicted by a model can be partitioned as $\boldsymbol{Z} = \{\boldsymbol{Z_o}, \boldsymbol{Z_u}\}$, where there are (noisy) observations $\boldsymbol{w}$ on $\boldsymbol{Z_o}$, but no observations of $\boldsymbol{Z_u}$. For example, we may have historic observations

on the output variable, and wish to predict future observations (forecasting), or we may have observations at a set of points in space and wish to predict values at locations in between (interpolation). Kennedy and O'Hagan (2001) propose a method for fully accounting for the uncertainty in $\boldsymbol{Z}$, given $\boldsymbol{w}$, via the model discrepancy within a Bayesian framework. However, in the context of health economics, we do not measure the costs and health consequences of sets of competing decisions, making this data driven method impossible. Specifying $p(\delta_z)$ directly therefore requires some form of elicitation of beliefs. See Garthwaite et al. (2005) and O'Hagan et al. (2006) for a discussion of methods.

Making meaningful judgements about the model discrepancy will be difficult, though it should always be possible to make a crude evaluation of a plausible range of orders of magnitude of δ_z, for example by asking questions like 'could the true incremental net benefit of decision 1 over decision 2 be a billion pounds greater than that predicted by the model, or a million pounds greater, or only a hundred pounds greater?' However, it may be easier to make judgements about δ_z indirectly. If we consider f in more detail we may be able to determine where in the model structural errors are likely to be located, and what their consequences might be. We therefore propose making judgements about discrepancy at the *sub-function* level.

3.2 Discrepancy at the 'sub-function' level

Any model f, except the most trivial, can be decomposed into a series of sub-functions that link the model inputs to the model output. So for example, a decomposition of the hypothetical model

$$Y = f(X_1, \ldots, X_7) = \left\{ (X_1 X_2 + X_3 X_4) \left(\frac{1}{1 + X_5} \right)^{-X_6} \right\} - X_7, \qquad (6)$$

might be in terms of sub-functions f_1, f_2 and f_3, with $Y_1 = f_1(X_1, \ldots, X_4) = X_1 X_2 + X_3 X_4$, $Y_2 = f_2(X_5, X_6) = \left(\frac{1}{1+X_5}\right)^{-X_6}$ and $Y = f_3(Y_1, Y_2, X_7) = Y_1 Y_2 - X_7$. The sub-functions f_1, f_2 and f_3 could be decomposed into further sub-functions, and so on. The inputs to each sub-function may contain both elements of the original input vector $\boldsymbol{X} = (X_1, \ldots, X_7)$ and outputs from other sub-functions in the decomposition. We call the output of each sub-function (unless it is the final model output, Y) an *intermediate parameter*.

For each sub-function, we ask the question 'would this sub-function, if evaluated at the true values of its inputs, result in the true value of the sub-function

output?' If not then we recognise potential structural error and introduce an uncertain discrepancy term, δ_i, either on the additive scale, i.e. $Y_i = f_i(\cdot) + \delta_i$, or multiplicative scale, i.e. $\log(Y_i) = \log\{f_i(\cdot)\} + \log(\delta_i)$. The idea is that, because each sub-function represents a much simpler process than the full model f, making judgements about discrepancy in f_i will be easier than making judgements about discrepancy in f.

Repeating the process for all sub-functions in the model will leave us with a series of n discrepancy terms, which we denote $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$. Note that for some sub-functions we will judge there is no structural error, usually when an intermediate parameter is by definition equal to the sub-function that generates it.

There will not usually be a unique decomposition of the model f into a series of sub-functions that links the model inputs $\boldsymbol{X}$ to the model output Y. However, some decompositions will be more useful than others for assessing discrepancy. Following the advice that it is preferable to elicit beliefs about observable quantities (O'Hagan et al., 2006), we search for decompositions where both inputs and outputs of the sub-functions are observable.

Once we have introduced discrepancy terms at the locations within the model where we judge there is potential structural error, we must make judgements about the discrepancies via the specification of the joint probability distribution $p(\boldsymbol{X}, \boldsymbol{\delta})$. We assume in our case study that discrepancies are independent of inputs, such that we can factorise the joint density $p(\boldsymbol{X}, \boldsymbol{\delta}) = p(\boldsymbol{X})p(\boldsymbol{\delta})$. This independence assumption does not need to hold for the discrepancy method to be valid, but specification of $p(\boldsymbol{\delta})$ independent of $p(\boldsymbol{X})$ will clearly be easier than specifying $p(\boldsymbol{X}, \boldsymbol{\delta})$.

We next consider the mean and variance for each discrepancy term δ_i, $i = 1, \ldots, n$. We make judgements about the sizes of the discrepancies relative to the mean values of the corresponding intermediate parameters, and set variances such that $\sqrt{\text{var}(\delta_i)} = v_i|E(Y_i)|$, with v_i chosen to reflect our judgements. Determining plausible values for v_i may not be a trivial task, a point to which we return in the discussion. We treat each δ_i as independent of all other uncertain quantities, unless there are constraints that prevent this (a constraint would arise, for example, in relation to a set of population proportion parameters that must sum to one) or unless there are good reasons to assume strong correlation between terms. Finally we select appropriate distributions with the specified means and variances.

Propagating the uncertainty we have specified for $\boldsymbol{\delta}$ through the model, along

with the uncertainty in the inputs, $\boldsymbol{X}$, allows us to check that the uncertainty in Z that our specification of $p(\boldsymbol{\delta})$ implies is plausible. If this is not the case then we must rethink our choice of distributions for the components of $\boldsymbol{\delta}$, most easily through altering our choices for v_i.

The sub-function discrepancy approach has two important consequences. Firstly, if we can adequately make judgements about all the discrepancy terms in the model (there may be many) then we will derive $p(\delta_z)$ and hence be able to make statements about our uncertainty about the incremental net benefit that incorporates beliefs about both inputs and structure. Perhaps more usefully though, we can use sensitivity analysis techniques to investigate the relative importance of the different structural errors, allowing us improve the parts of the model where this is most needed. If, after repeating the sensitivity analysis in our improved model, we find that discrepancies now have a lesser impact on the output uncertainty, then we have in an important sense built a more robust model structure.

4 Applying the sub-function discrepancy method to our physical activity model

We return to our base case physical activity model, and beginning at the net benefit equation (3), work 'backwards' through the model, assessing potential structural error at each sub-function.

4.1 Assessment of sub-function generating the output parameter Y

The model output, Y predicts the incremental net benefit, as defined in equation (3). Evaluation of equation (3) at the true values of ΔQ and ΔC would, by definition, result in the true value of the incremental net benefit, Z, so there is no structural error at this point in the model, and therefore no discrepancy term.

4.2 Assessment of sub-function generating the intermediate parameter ΔQ

The incremental health effect of the intervention over the non-intervention, ΔQ is

$$\Delta Q = \sum_{j=1}^{8} p_{2j} q_{2j} - \sum_{j=1}^{8} p_{1j} q_{1j}, \tag{7}$$

where p_{dj} and q_{dj} are the probabilities and discounted health effects in QALYs respectively for health state h_j under decision d. Future health effects (and future costs) are discounted to reflect time preference whereby higher value is placed on benefits that occur in the near future than on those occurring in the distant future. See Krahn and Gafni (1993) for a discussion of the role of discounting in health economic evaluation.

Health effects for each state are assumed to be equal regardless of the decision, i.e. that $q_{1j} = q_{2j} = q_j$, and therefore that

$$\Delta Q = \sum_{j=1}^{8} (p_{2j} - p_{1j}) q_j = \sum_{j=1}^{8} (p_{2j} - p_{1j})(q_j - q_1) = \sum_{j=1}^{8} (p_{2j} - p_{1j}) q_j^{(dec)}, \tag{8}$$

where the final term is a re-expression in terms of the *decremental* health effect, $q_j^{(dec)}$ relative to the disease free state $j = 1$.

We ask the question, 'given the true values of p_{dj} and q_j, does (8) result in the true value of ΔQ?' Because we imagine that the intervention could have an impact on a number of diseases other than CHD, stroke and diabetes we recognise potential structural error and introduce an uncertain additive discrepancy term, $\delta_{\Delta Q}$ into (8), which becomes

$$\Delta Q = \sum_{j=1}^{8} (p_{2j} - p_{1j}) q_j^{(dec)} + \delta_{\Delta Q}. \tag{9}$$

Since exercise can result in poor health outcomes as well as good outcomes, for example through musculo-skeletal injuries or accidents, we specify a mean of zero for $\delta_{\Delta Q}$. We could assume a non-zero mean for $\delta_{\Delta Q}$ if we felt that increased exercise was likely to be on balance beneficial. This will have the effect of shifting the mean of the model output unless the sub-function related to the discrepancy is entirely unimportant. Introducing discrepancy terms that have non-zero mean may well be reasonable, but by doing so we are effectively making a judgement that the base case model is 'wrong'.

We judge that $\delta_{\Delta Q}$ is unlikely to be more than $\pm 10\%$ of ΔQ, and we represent our beliefs about $\delta_{\Delta Q}$ using a normal distribution with a standard deviation equal to 5% of the mean of ΔQ, i.e. $\delta_{\Delta Q} \sim \mathrm{N}[0, \{0.05 \times E(\Delta Q)\}^2]$.

4.3 Assessment of sub-function generating the intermediate parameter ΔC

The incremental cost of the intervention over the non-intervention, ΔC is

$$\Delta C = \sum_{j=1}^{8} p_{2j} c_{2j} - \sum_{j=1}^{8} p_{1j} c_{1j}, \tag{10}$$

where p_{dj} and c_{dj} are the probabilities and discounted costs respectively that are associated with health state h_j under decision d.

Costs, not including the cost of the intervention itself c_0, are assumed to be equal across decision arms, i.e. that $c_{2j} = c_{1j} + c_0$, and therefore that

$$\Delta C = \sum_{j=1}^{8} p_{2j}(c_{1j} + c_0) - \sum_{j=1}^{8} p_{1j} c_{1j} = c_0 + \sum_{j=1}^{8} (p_{2j} - p_{1j}) c_{1j}, \tag{11}$$

where c_0 is a model input.

As above, there may be costs that relate to diseases other than CHD, stroke and diabetes that are not included in ΔC and we therefore introduce an additive discrepancy term, $\delta_{\Delta C}$, and specify that $\delta_{\Delta C} \sim \mathrm{N}[0, \{0.05 \times E(\Delta C)\}^2]$.

4.4 Assessment of sub-function generating the intermediate parameters c_{1j}

The intermediate parameters c_{1j} represent the discounted cost associated with the eight health states. In the base case model the costs for the eight states are derived from the costs associated with the three individual diseases, with costs for comorbid states assumed to be the sum of the costs for the constituent diseases, so for example

$$c_{1,8} = c_{chd} + c_{str} + c_{dm}. \tag{12}$$

Costs may not be additive in this way, so we introduce additive discrepancy terms, δ_{c_j}, for the intermediate parameters that relate to the comorbid states, c_{1j} $j = 5, \ldots, 8$.

We judge that comorbid state costs could be higher or lower than the sum of the constituent costs, so we assumed a mean of zero for each discrepancy term, $\delta_{c_j},\ j = 5,\ldots,8$. We represent beliefs about δ_{c_j} via $\delta_{c_j} \sim \mathrm{N}[0, \{0.05 \times E(c_j)\}^2],\ j = 5,\ldots,8$.

4.5 Assessment of sub-function generating the intermediate parameters c_{chd}, c_{str} and c_{dm}

The discounted costs for CHD, stroke and diabetes are

$$c_k = c_k^* \times \alpha_k, \tag{13}$$

where k indexes the set $\{CHD, stroke, diabetes\}$. Costs (other than the cost of the intervention) are assumed to occur at some time in the future, and are discounted at 3.5% per year. The parameters c_k^* represent undiscounted costs, and α_k, are the discounting factors for the length of time between the intervention and the occurrence of the relevant health outcomes.

Given true values for c_k^* and α_k equation (13) will result in a true value for c_k, and there is no structural error at this point.

4.6 Assessment of sub-function generating the intermediate parameters c^*_{chd}, c^*_{str} and c^*_{dm}

The undiscounted mean per-person lifetime costs for CHD, stroke and diabetes are

$$c_k^* = \frac{t_k}{n_k}\left(age_k^{(dth)} - age_k^{(onst)}\right), \tag{14}$$

where k indexes the set $\{CHD,\ stroke,\ diabetes\}$, and where t_k are total annual NHS costs for disease k, and where n_k are UK prevalent cases of disease k for the same year. The parameters t_k, n_k, $age_k^{(dth)}$ and $age^{(onst)}$ are model inputs.

Mean per person undiscounted costs are calculated as the mean per person annual NHS cost multiplied by the mean length of time in the disease state. If the per person per year cost of disease is dependent on the length of time the individual spends in the disease state (e.g. if costs are greater near to the end of life), then c^*_{chd}, c^*_{str} and c^*_{dm} as calculated will not equal the mean per person per year costs. To properly calculate the mean we need to know the joint distribution of the costs and length of time in the disease state. To account for the difference we introduce discrepancy terms $\delta_{c_k^*}$.

We judge that disease costs could in reality be higher or lower than the modelled costs as a result of the structural error, so we assume a mean of zero for each discrepancy term, $\delta_{c_k^*}$. We represent beliefs about $\delta_{c_k^*}$ via $\delta_{c_k^*} \sim \text{N}[0, \{0.05 \times E(c_k^*)\}^2]$.

4.7 Assessment of sub-function generating the intermediate parameters α_{chd}, α_{str} and α_{dm}

The discounting factors for CHD, stroke and diabetes are

$$\alpha_k = \left(\frac{1}{1+\theta}\right)^{l_k}, \tag{15}$$

where l_k is the mean length of life remaining at the time of intervention for disease $k \in \{CHD, \textit{stroke}, \textit{diabetes}\}$, and θ is the per-year discount rate for both costs and health effects. The mean length of life remaining, l_k, is given by

$$l_k = \frac{1}{2}\left(age_k^{(onst)} + age_k^{(dth)}\right) - age^{(int)}, \tag{16}$$

where $age_k^{(onst)}$ is the mean age of onset of disease k, $age_k^{(dth)}$ is the mean age of death from disease k and $age^{(int)}$ is the mean age of the cohort at the time of the intervention. The parameters θ, $age_k^{(dth)}$, $age^{(onst)}$ and $age^{(int)}$ are model inputs.

In the base case model we assume that the costs of each disease will be realised at a time midway between the average age of disease onset, and the average age of death from that disease. This is not necessarily true and we introduce additive discrepancy terms δ_{α_k}.

Discount factors must lie in $(0, 1]$, and so discrepancies must lie in $(-\alpha_k, 1-\alpha_k]$. To satisfy this constraint we assume that $\alpha_k + \delta_{\alpha_k}$ follows a beta distribution. We have no reason to believe that the true values of the discount rates will be higher or lower than the modelled values, so we assume that δ_{α_k} has mean zero for all k. As above, we assume that the standard deviation is 5% of the mean value of the intermediate parameter, i.e. that $\sqrt{\text{var}(\delta_{\alpha_k})} = 0.05E(\alpha_k)$.

See Appendix C for details of the calculation of Dirichlet distribution hyperparameters that satisfy these requirements. The more general Dirichlet distribution specification of uncertainty is required for other discrepancy terms in the model, so for brevity we treat $\alpha_k + \delta_{\alpha_k}$ and $1 - (\alpha_k + \delta_{\alpha_k})$ as 'sum-to-one' parameters and the beta distribution as a special case of the Dirichlet distribution.

4.8 Assessment of sub-function generating the intermediate parameters $q_j^{(dec)}$

The intermediate parameters $q_j^{(dec)}$ represent the discounted decremental health effects (in QALYs) associated with the eight health states. In the base case model these terms are derived from the discounted decremental health effects associated with the three individual diseases, with decremental effects for comorbid states assumed to be the sum of the decremental effects for the constituent diseases.

This means that, for example

$$q_8^{(dec)} = q_{chd}^{(dec)} + q_{str}^{(dec)} + q_{dm}^{(dec)}, \tag{17}$$

where the parameters $q_{chd}^{(dec)}$, $q_{str}^{(dec)}$ and $q_{dm}^{(dec)}$ are model inputs. Decremental health effects may not be additive in this way, so we introduce discrepancy terms δ_{q_j} for the comorbid health states $j = 5, \ldots, 8$.

We judge that comorbid state decremental health effects could be higher or lower than the sum of the constituent terms, so assume a mean of zero for each discrepancy term, δ_{q_j}, $j = 5, \ldots, 8$. We represent beliefs about δ_{q_j} via $\delta_{q_j} \sim \text{N}[0, \{0.05 \times E(q_j)\}^2]$, $j = 5, \ldots, 8$.

4.9 Assessment of sub-function generating the intermediate parameters p_{dj}

The proportions of the population who are expected to experience each disease state $j = 1, \ldots, 8$ under decision options $d = 1, 2$ are

$$p_{dj} = p_d^{(ex)}\, p_d^{(mnt)}\, r_j^{(ex)} + p_d^{(ex)} \left(1 - p_d^{(mnt)}\right) r_j^{(sed)} + \left(1 - p_d^{(ex)}\right) r_j^{(sed)}, \tag{18}$$

where $r_j^{(ex)}$ and $r_j^{(sed)}$ are the risks of disease state j in those who exercise and in those who are sedentary, respectively. The probability of new exercise under decision option d is $p_d^{(ex)}$, and the probability of maintenance of exercise is $p_d^{(mnt)}$. The parameters $p_d^{(ex)}$ and $p_d^{(mnt)}$ are model inputs.

Parameters defining health state probabilities lie in $[0, 1]$, and must sum to one over j, so discrepancies must lie in $[-p_{dj}, 1 - p_{dj}]$, and must sum to zero over j. To satisfy this constraint we assume a Dirichlet distribution for $p_{dj} + \delta_{p_{dj}}$.

We have no reason to believe that the true values of the health state probabilities would be higher or lower than the modelled values, so we assume that

$E(\delta_{p_{dj}}) = 0,\ \forall d, j$. We assume that the standard deviation was 5% of the mean value of the intermediate parameter, i.e.

$$\frac{1}{8}\sum_{j=1}^{8}\frac{\sqrt{\text{var}(\delta_{p_{dj}})}}{E(p_{dj})} = 0.05. \tag{19}$$

See Appendix C for details of the calculation of the Dirichlet hyperparameters that satisfy these requirements.

4.10 Assessment of sub-function generating the intermediate parameters $r_j^{(ex)}$ and $r_j^{(sed)}$

The parameters $r_j^{(ex)}$ and $r_j^{(sed)}$ represent the risks of health state j in a population that exercises and in a sedentary population, respectively. In the base case model we assume that occurrences of CHD, stroke and diabetes are independent, and therefore that the $r_{chd}^{(ex)}$, $r_{str}^{(ex)}$ and $r_{dm}^{(ex)}$ act multiplicatively to generate the $r_j^{(ex)}$ (and similarly multiplicatively in the sedentary population). So for example,

$$r_1^{(ex)} = (1 - r_{chd}^{(ex)})(1 - r_{str}^{(ex)})(1 - r_{dm}^{(ex)}). \tag{20}$$

We assume that occurrences of CHD, stroke and diabetes are independent, which may not be true, so we introduce additive discrepancy terms $\delta_{r_j^{(sed)}}$ and $\delta_{r_j^{(ex)}}$. Following the same argument as that in 4.9 we assume a Dirichlet distributions for $r_j^{(ex)} + \delta_{r_j^{(ex)}}$ and for $r_j^{(sed)} + \delta_{r_j^{(sed)}}$. We have no reason to believe that the true values of the disease risks would be higher or lower than the modelled values, so we assume that $E(\delta_{r_j^{(ex)}}) = E(\delta_{r_j^{(sed)}}) = 0,\ \forall j$. We assume that the standard deviations were 5% of the mean values of the intermediate parameters, i.e.

$$\frac{1}{8}\sum_{j=1}^{8}\frac{\sqrt{\text{var}\left(\delta_{r_j^{(ex)}}\right)}}{E\left(r_j^{(ex)}\right)} = \frac{1}{8}\sum_{j=1}^{8}\frac{\sqrt{\text{var}\left(\delta_{r_j^{(sed)}}\right)}}{E\left(r_j^{(sed)}\right)} = 0.05. \tag{21}$$

4.11 Assessment of sub-function generating the intermediate parameters $r_k^{(ex)}$

The parameters $r_k^{(ex)}$ where k indexes the set $\{CHD, stroke, diabetes\}$ represent the risks of CHD, stroke and diabetes in those who exercise. They are calculated by multiplying baseline risk by the relative risk of disease given exercise, i.e.

$$r_k^{(ex)} = r_k^{(sed)} \times RR_k, \tag{22}$$

where $r_k^{(sed)}$ and RR_k are model inputs.

Given true values for $r_k^{(sed)}$ and RR_k, sub-function (22) will result in the true value of $r_k^{(ex)}$ by definition of a relative risk, so there is no structural error at this point.

5 Results of discrepancy analysis

A total of 48 discrepancy terms were introduced into the model. The addition of the discrepancy terms 'corrects' any structural error, and allows us now to write

$$Z = f^*(\boldsymbol{X}, \boldsymbol{\delta}), \tag{23}$$

where f^* takes the same functional form as f, but with the inclusion of the discrepancy terms as described in section 4.

5.1 Model output after inclusion of discrepancy terms

We sampled the input and discrepancy distributions and ran the model f^* 100,000 times. This resulted in a predicted mean incremental net benefit of £247, which is equal to the that predicted by the base case model. The 95% credible interval was -£886 to £1444, which is wider than that of the base case model, reflecting the recognition of our additional uncertainty about the true incremental net benefit due to possible model structural error.

Returning to figure 2, we note the larger cloud of points on the cost-effectiveness plane (figures 2a and 2b), reflecting the additional uncertainty. The additional uncertainty has reduced the probability that the intervention is cost-effective, $P(\text{INB} > 0)$, at $\lambda =$£20,000 to 0.66 (closer to the value of 0.5 that represents complete uncertainty), and flattened the cost effectiveness acceptability curve towards the horizontal line at $P(\text{INB} > 0) = 0.5$ (figure 2c). The additional uncertainty is also reflected in the wider empirical distribution in figure 2d.

5.2 Determining important structural errors via variance based sensitivity analysis

Following our analysis of structural error we may then wish to make improvements to the model. It is unlikely that all the sub-model discrepancy terms are equally 'important', by which we mean that some terms may be located in parts of the

model in which structural errors contribute very little to uncertainty about Z, the incremental net benefit. If we can identify the most important discrepancy terms, we can consider reducing structural errors through better modelling, perhaps by relaxing certain assumptions, or by including features that were omitted initially. Similarly, identifying unimportant discrepancy terms will tell us where it is *not* worth improving the model.

Note that any re-modelling following a sensitivity analysis may not reduce uncertainty about Z, for example if the improved model structure introduces new, uncertain parameters. In this situation we are effectively 'transferring' our uncertainty from structure to inputs. This may be helpful simply because input uncertainty is generally easier to manage, but in any case we believe that a formal consideration of the balance between uncertainty due to model structure and uncertainty due to model inputs is desirable.

We can identify a set of important discrepancy terms using standard sensitivity analysis techniques. A considerable number of methods exist (Saltelli et al., 2008), but for our purposes we have chosen to use a variance based sensitivity analysis approach. In this approach the measure of importance for each discrepancy term, δ_i $i = 1, \ldots, n$, is defined as its 'main effect index',

$$\frac{\mathrm{var}_{\delta_i}\{E(Z|\delta_i)\}}{\mathrm{var}(Z)}. \tag{24}$$

Given the identity $\mathrm{var}(Z) = \mathrm{var}_{\delta_i}\{E(Z|\delta_i)\} + E_{\delta_i}\{\mathrm{var}(Z|\delta_i)\}$ the main effect index gives the expected reduction in the variance of Z obtained by learning the value of δ_i.

The main effect index for uncorrelated discrepancy terms is straightforward to calculate using Monte Carlo methods. In this case $E(Z|\delta_i)$ can be approximated by

$$E(Z|\delta_i) \approx \frac{1}{S}\sum_{s=1}^{S} f^*(\boldsymbol{x}_s, \boldsymbol{\delta}_{-i,s}, \delta_i), \tag{25}$$

where $\{(\boldsymbol{x}_s, \boldsymbol{\delta}_{-i,s}), s = 1, \ldots, S\}$ is a (large) sample from the distribution $p(\boldsymbol{X}, \boldsymbol{\delta}_{-i})$.

However, if δ_i is correlated with other discrepancy terms or inputs, then this method would require us to draw samples from the conditional distribution $p(\boldsymbol{X}, \boldsymbol{\delta}_{-i}|\delta_i)$. Such conditional distributions may not be known, so we propose an alternative approximation method. See appendix B.

Following a variance based sensitivity analysis of the discrepancy terms in our model, eight of the terms appeared to be important, having main effects $> 5\%$.

The pattern of importance suggests that re-expressing the sub-functions for the parameters p_{dj} is key to reducing structural error (table 1).

Table 1: Main effect indexes for discrepancy terms ($> \mathbf{5}\%$ in bold)

Discrepancy	Main effect	Discrepancy	Main effect	Discrepancy	Main effect
$\delta_{r_1^{(ex)}}$	0.002	$\delta_{p_{1,1}}$	**0.266**	$\delta_{c^*_{chd}}$	0.002
$\delta_{r_2^{(ex)}}$	0.002	$\delta_{p_{1,2}}$	**0.128**	$\delta_{c^*_{str}}$	0.002
$\delta_{r_3^{(ex)}}$	0.003	$\delta_{p_{1,3}}$	**0.076**	$\delta_{c^*_{dm}}$	0.001
$\delta_{r_4^{(ex)}}$	0.002	$\delta_{p_{1,4}}$	0.002	$\delta_{d_{chd}}$	0.002
$\delta_{r_5^{(ex)}}$	0.003	$\delta_{p_{1,5}}$	**0.054**	$\delta_{d_{str}}$	0.002
$\delta_{r_6^{(ex)}}$	0.002	$\delta_{p_{1,6}}$	0.025	$\delta_{d_{dm}}$	0.002
$\delta_{r_7^{(ex)}}$	0.003	$\delta_{p_{1,7}}$	0.014	δ_{q_5}	0.002
$\delta_{r_8^{(ex)}}$	0.004	$\delta_{p_{1,8}}$	0.010	δ_{q_6}	0.002
$\delta_{r_1^{(sed)}}$	0.002	$\delta_{p_{2,1}}$	**0.257**	δ_{q_7}	0.002
$\delta_{r_2^{(sed)}}$	0.002	$\delta_{p_{2,2}}$	**0.124**	δ_{q_8}	0.002
$\delta_{r_3^{(sed)}}$	0.002	$\delta_{p_{2,3}}$	**0.076**	δ_{c_5}	0.002
$\delta_{r_4^{(sed)}}$	0.002	$\delta_{p_{2,4}}$	0.002	δ_{c_6}	0.002
$\delta_{r_5^{(sed)}}$	0.002	$\delta_{p_{2,5}}$	0.049	δ_{c_7}	0.002
$\delta_{r_6^{(sed)}}$	0.002	$\delta_{p_{2,6}}$	0.025	δ_{c_8}	0.002
$\delta_{r_7^{(sed)}}$	0.002	$\delta_{p_{2,7}}$	0.013	δ_{Δ_q}	0.003
$\delta_{r_8^{(sed)}}$	0.002	$\delta_{p_{2,8}}$	0.008	δ_{Δ_c}	0.001

5.3 The relative importance of parameter to structural error uncertainty

We may also wish to understand the relative importance of the contributions of uncertainty about structural error and uncertainty about input parameters to the overall uncertainty in Z. We can measure this using the *structural parameter uncertainty ratio*, which we define as

$$\frac{\text{var}_{\boldsymbol{\delta}}\{E_{\boldsymbol{X}}(Z|\boldsymbol{\delta})\}}{\text{var}_{\boldsymbol{X}}\{E_{\boldsymbol{\delta}}(Z|\boldsymbol{X})\}}. \tag{26}$$

This is straightforward to calculate if $\boldsymbol{\delta}$ is independent of $\boldsymbol{X}$ since $E_{\boldsymbol{X}}(Z|\boldsymbol{\delta} = \boldsymbol{\delta}') = E_{\boldsymbol{X}}\{f^*(\boldsymbol{X}, \boldsymbol{\delta})|\boldsymbol{\delta} = \boldsymbol{\delta}'\} = E_{\boldsymbol{X}}\{f^*(\boldsymbol{X}, \boldsymbol{\delta}')\}$ and $E_{\boldsymbol{\delta}}(Z|\boldsymbol{X} = \boldsymbol{x}) = E_{\boldsymbol{\delta}}\{f^*(\boldsymbol{X}, \boldsymbol{\delta})|\boldsymbol{X} =$

$\boldsymbol{x}\} = E_{\boldsymbol{\delta}}\{f^*(\boldsymbol{x}, \boldsymbol{\delta})\}$. If $\boldsymbol{\delta}$ and $\boldsymbol{X}$ are not independent calculating the conditional expectations is more difficult, though methods are available (Oakley and O'Hagan, 2004).

The structural parameter uncertainty ratio in our model is 2.0 indicating that, given our specification of discrepancy, learning the discrepancy terms would result in double the expected reduction in the variance of the output compared with the expected reduction in variance on learning the true values of all the input parameters.

5.4 Analysis of robustness to different choices of v_i

In our case study we set v_i (the ratio of the discrepancy standard deviation to the mean of the corresponding intermediate parameter) to 5% equally for all discrepancy terms, judging this to be an appropriate reflection of the likely range of structural error. The resulting additional uncertainty in the model output was plausible, and the variance based sensitivity analysis implied that there was important structural error in the sub-model that generates the health state probability parameters, p_{dj} (section 4.9).

In order to test the robustness of our conclusion to minor variations in the specification of the discrepancies we altered values for v_i over a plausible range. We grouped the discrepancy terms into four sets: terms relating to cost parameters, terms relating to health effect parameters, terms relating to population proportion parameters, and terms relating to the discount factors. Within each set the values for v_i were either doubled, halved or maintained at 5%. Given three levels for v_i and four sets of discrepancy terms there are $3^4 = 81$ combinations of choices for v_i including our original specification of $v_i = 5\%$ for all i.

In all 81 cases a very similar pattern of main effect indexes to that reported in table 1 was observed, with the $\delta_{p_{dj}}$ terms dominating, indicating robustness to choices of v_i over the range 2.5% to 10%.

5.5 Remodelling the sub-functions where there is important structural error

Variance based sensitivity analysis has identified $\delta_{p_{dj}}$ to be important discrepancy terms, indicating that we have important structural error in the sub-model that generates the health state probability parameters, p_{dj}.

In the base case model the proportion of people who begin and then maintain exercise is assumed constant over time. If we believe that there will be a decline in the proportion of people who exercise over time then we could re-structure the model sub-function to reflect this. We could, for example, assume an exponential decline, whereby the proportion exercising at each year in the future is equal to the proportion exercising in the previous year multiplied by some (uncertain) constant. If the risk of each disease state j decreased (increased for the well state) linearly from $r_j^{(sed)}$ to $r_j^{(ex)}$ with increasing time spent exercising (with a threshold achieved after, say, four years exercise), then we could write

$$\begin{aligned} p_{dj} &= \left(1 - p_d^{(ex)}\right) r_j^{(sed)} + p_d^{(ex)} \left(1 - m_d\right) r_j^{(sed)} \\ &+ p_d^{(ex)} \left(m_d - m_d^2\right) \left(\frac{1}{4} r_j^{(ex)} + \frac{3}{4} r_j^{(sed)}\right) + p_d^{(ex)} \left(m_d^2 - m_d^3\right) \left(\frac{1}{2} r_j^{(ex)} + \frac{1}{2} r_j^{(sed)}\right) \\ &+ p_d^{(ex)} \left(m_d^3 - m_d^4\right) \left(\frac{3}{4} r_j^{(ex)} + \frac{1}{4} r_j^{(sed)}\right) + p_d^{(ex)} m_d^4 r_j^{(ex)}, \end{aligned} \tag{27}$$

where m_d is the proportion of the population who exercised in year t who continue to exercise in year $t+1$, under decision d.

To complete the new model specification we need to specify distributions for m_1 and m_2. We assume that m_1 and m_2 are jointly normally distributed with means of 0.5, variances of 0.01 and a correlation of 0.9.

5.6 Results following sub-function remodelling

The mean net benefit following remodelling was £71 (-£273 to £572), with the probability that the intervention is cost-effective, $P(\text{INB} > 0)$, at λ =£20,000 equal to 0.59. Returning again to figure 2 we see that there is now a smaller cloud a points on the cost-effectiveness plane, and that these are shifted towards the left and the line of no effect (at $\Delta Q = 0$). The cost-effectiveness acceptability curve (figure 2c) suggests that following remodelling we predict that the intervention has a lower probability of being cost-effective than predicted by the base case model at all values of λ. The leftwards shift of the incremental net benefit density towards zero supports this (figure 2d).

By re-structuring the important sub-function in the model to better incorporate our beliefs about real-world processes, we find that the incremental net benefit distribution is shifted downwards. This is due to our judgement that a proportion of those who begin new exercise will cease exercising, and that instead of this drop being a single step change, the fall will be exponential over time.

This results in a lower proportion of maintained exercise in both the intervention and non-intervention groups, and a lower absolute reduction in disease risk and smaller incremental benefit.

6 Discussion

We have presented a discrepancy modelling approach that allows us to quantify our judgements about how close model predictions will be to reality. We incorporate our beliefs about structural error through the addition of discrepancy terms at the sub-function level throughout the model, and following this we are able to determine the sources of structural error that have an important impact on the output uncertainty. Without the model decomposition and variance based sensitivity analysis it may not be at all obvious which are the most important sources of structural error, and so the method reveals features of the model that are otherwise hidden.

As is clear from our description of the model in section 2.1, a model's structure rests upon a series of assumptions regarding the relationships between the inputs, the intermediate parameters and the output. In any modelling process it is unavoidable that such assumptions are made, and in one sense model building is just a formal representation of a set of assumptions in mathematical functional form. Health economic modellers sometimes explore the sensitivity of the model prediction to underlying assumptions in a "what if" scenario analysis in which sets of alternative assumptions are modelled (see Bojke et al. (2009) for a review of the methods that are currently used to manage health economic evaluation model uncertainty, and Kim et al. (2010) for a specific example of modelling alternative scenarios). However, this process cannot in any formal sense quantify the sensitivity of the results to the assumptions, and nor can it quantify any resulting prediction uncertainty. Our method is an attempt to formally quantify the effect of all assumptions in the model about which we do not have complete certainty.

The method is most useful as a sensitivity analysis tool, highlighting areas of the model that may require further thought. However, if the modeller can satisfactorily specify a joint distribution for the inputs and the discrepancies, then the method results in a proper quantification of uncertainty about the 'true' incremental net benefit of one decision over an alternative, taking into account judgements about both parameters and structure.

6.1 Model complexity and parsimony

Current good practice guidance on modelling for health economic evaluation states that a model should only be as complex as necessary (Weinstein et al., 2003), but this well intentioned advice does not actually help us make judgements about how complex any particular model should be. Another guiding principle is the requirement for a model to be comprehensible to the non-modeller: a decision maker's trust in a model can easily be eroded if the model is so complicated that its features cannot be easily communicated (Taylor-Robinson et al., 2008).

Our view is that, in the health economic context, increasing the model complexity has the effect of transferring uncertainty about structural error, which we express through the specification of model discrepancy terms, to uncertainty about model input parameters. Structural error arises when a simple model is used to a model a complex real world process, thereby omitting aspects that could effect costs or consequences. If we make the model more complex by including such omitted features, typically we will then have more input parameters in the model.

Increasing the complexity of a model will therefore be desirable if the additional complexity relates to parts of the model in which discrepancy terms are influential, and if we have suitable data to tell us about any extra parameters that are required. This is because, to the decision-maker, data-driven probability distributions for model parameters will be preferable to distributions on (plausibly large) discrepancy terms based solely on subjective judgements of the modeller.

Our framework can help guide the choice of model complexity by identifying which discrepancy terms are likely to be important. If we are satisfied that a structural error will have little effect on the model output, then increasing the complexity of the model to reduce such an error is likely to have little benefit.

6.2 Extension to a scenario with more than two decision options

In our case study there were two competing decisions, and therefore a single obvious scalar model output quantity: the incremental net benefit. This allowed a straightforward analysis of sub-function discrepancy importance using the variance based sensitivity method. However, when there are more than two competing decisions there is no single, scalar model output that is equivalent of the incremental net benefit, and therefore it is not immediately obvious how to proceed with variance based sensitivity methods. One solution would be to work instead within

an expected value of information framework, defining important model sub-unit discrepancy terms as those which have an expected value above some threshold.

6.3 How might this work in practice?

We envisage that the sub-function discrepancy approach has the greatest potential if used prospectively during model building. This will allow the modeller to incorporate judgements about structural error as they construct the model, encouraging an explicit recognition of the potential impact of the structural choices.

Model development is a sequential, hierarchical, iterative process of uncovering and evaluating options regarding structure, parameterisation and incorporation of evidence (Chilcott et al., 2010a). The process depends on the modeller developing an understanding of the decision problem, which is by its nature subjective. This understanding of the decision problem is the foundation upon which judgements are made in the model building process, and also provides the basis for making judgements about the likely discrepancy inherent in different model formulations. The essence of the discrepancy approach is that it allows a *formal quantification* of the impact of the choices made throughout the model building process.

Ultimately, the validity of the method relies on the ability to meaningfully specify the joint distribution of inputs and discrepancies, $p(\boldsymbol{X}, \boldsymbol{\delta})$. In our study we represented our beliefs about $p(\boldsymbol{X}, \boldsymbol{\delta})$ fairly crudely, making assumptions of independence between inputs and discrepancies and independence between groups of discrepancies that were not otherwise constrained. Key to the specification of the discrepancy in our case study was the choice of values for v_i that control the variance of δ_i relative to the mean of the corresponding intermediate parameter. We determined a value for each v_i by informally eliciting our own judgements about the plausible range for the structural error relative to the size of the intermediate parameter. We then examined the effect of making different sets of choices in a sensitivity analysis.

Whilst we felt that this was sufficient in our case study for the purposes of identifying important model sub-functions we recognise that making defensible judgements about model discrepancies is in general likely to be difficult. If we wish to proceed to a full quantification of our uncertainty about the target quantity via $Z = f^*(\boldsymbol{X}, \boldsymbol{\delta})$ then a more sophisticated specification of $p(\boldsymbol{X}, \boldsymbol{\delta})$ will typically be required. Developing practical methods for making helpful judgements about $p(\boldsymbol{X}, \boldsymbol{\delta})$ is an area for future research.

Acknowledgements

MS is funded by UK Medical Research Council fellowship grant G0601721.

Appendix A - Base case model input parameters

Table 2: Uncertain inputs and their distributions

Input	Label	Description	Distribution	Hyperparameters	
X_1	c_0	Intervention cost (£)	gamma	shape=100; scale=1	
X_2	t_{chd}	Total NHS costs (2005) for CHD (£)	gamma	sh=3.677×10^9; sc=1	
X_3	t_{str}	Total NHS costs (2005) for stroke (£)	gamma	sh=2.872×10^9; sc=1	
X_4	t_{dm}	Total NHS costs (2005) for diabetes (£)	gamma	sh=5.314×10^9; sc=1	
X_5	n_{chd}	Number of UK cases of CHD	Poisson	$\mu = 2.60 \times 10^6$	
X_6	n_{str}	Number of UK cases of stroke	Poisson	$\mu = 1.40 \times 10^6$	
X_7	n_{dm}	Number of UK cases of diabetes	Poisson	$\mu = 1.53 \times 10^6$	
X_8	$q_{chd}^{(dec)}$	Discounted decremental health effect for CHD (QALYs)	normal	$\mu = 6.71$; $\sigma = 0.048$	
X_9	$q_{str}^{(dec)}$	Discounted decremental health effect for stroke (QALYs)	normal	$\mu = 10.23$; $\sigma = 0.048$	
X_{10}	$q_{dm}^{(dec)}$	Discounted decremental health effect for DM (QALYs)	normal	$\mu = 2.08$; $\sigma = 0.048$	
X_{11}	$p_1^{(ex)}$	Probability of new exercise - non-intervention group	MVN	$\mu = 0.246$; $\sigma = 0.038$	$\rho = 0.5$
X_{12}	$p_2^{(ex)}$	Probability of new exercise - intervention group		$\mu = 0.294$; $\sigma = 0.040$	
X_{13}	$p_1^{(mnt)}$	Probability exercise is maintained - non-intervention	MVN	$\mu = 0.5$; $\sigma = 0.1$	$\rho = 0.9$
X_{14}	$p_2^{(mnt)}$	Probability exercise is maintained - intervention		$\mu = 0.5$; $\sigma = 0.1$	
X_{15}	$r_{chd}^{(sed)}$	Risk of CHD in a sedentary group	beta	$\alpha = 80$; $\beta = 385$	
X_{16}	$r_{str}^{(sed)}$	Risk of stroke in a sedentary group	beta	$\alpha = 226$; $\beta = 4072$	
X_{17}	$r_{dm}^{(sed)}$	Risk of diabetes in a sedentary group	beta	$\alpha = 346$; $\beta = 3344$	
X_{18}	RR_{chd}	Relative risk of CHD in active vs sedentary pop	lognormal	$\mu = 0.666$; $\sigma = 0.130$	
X_{19}	RR_{str}	Relative risk of stroke in active vs sedentary pop	lognormal	$\mu = 0.720$; $\sigma = 0.343$	
X_{20}	RR_{dm}	Relative risk of diabetes in active vs sedentary pop	lognormal	$\mu = 0.710$; $\sigma = 0.123$	
X_{21}	$age^{(onst)}$	Average age of onset of disease (same for all diseases)	normal	$\mu = 57.5$; $\sigma = 2$	
X_{22}	$age_{chd}^{(dth)}$	Average age of death for CHD (years)	normal	$\mu = 71$; $\sigma = 2$	
X_{23}	$age_{str}^{(dth)}$	Average age of death for stroke (years)	normal	$\mu = 59$; $\sigma = 2$	
X_{24}	$age_{dm}^{(dth)}$	Average age of death for diabetes (years)	normal	$\mu = 61$; $\sigma = 2$	

Table 3: Fixed inputs

Input	Label	Description	Value
X_{25}	$age^{(int)}$	Average age of cohort at time of intervention (years)	50
X_{26}	θ	Discount rate (per year)	0.035
X_{27}	λ	Willingness to pay (£/QALY)	20,000

Appendix B - Algorithm for calculating main effect index when model inputs are correlated

We have a model $y = f(\boldsymbol{x})$ with p inputs, $\boldsymbol{x} = \{x_1, \ldots, x_p\}$ and a scalar output y. We are uncertain about the input values, and therefore write $\boldsymbol{X}$ and represent beliefs via $p(\boldsymbol{X})$. Note that to use this method to determine the main effect indexes for the discrepancy terms, $\boldsymbol{\delta}$, we treat the discrepancies as just another set of uncertain model inputs, so in the description below, $\boldsymbol{\delta}$ would be included in the vector of all uncertain input quantities, $\boldsymbol{X}$.

We are interested in the sensitivity of the model output to the p model inputs and measure this using the 'main effect index', defined for input X_i as

$$\frac{\text{var}_{X_i}\{E_{X_{-i}}(Y|X_i)\}}{\text{var}(Y)}, \tag{28}$$

where $\boldsymbol{X}_{-i} = \{X_1, \ldots, X_{i-1}, X_{i+1} \ldots, X_p\}$.

At first sight this is non-trivial via Monte Carlo methods if X_i is not independent of $\boldsymbol{X}_{-i}$ since calculating $E_{X_{-i}}(Y|X_i)$ requires sampling from the conditional distribution $\boldsymbol{X}_{-i}|X_i$, which may not be explicitly known. We therefore suggest the following alternative method, which does not require us to sample from the conditional distributions.

We first obtain a single Monte Carlo sample $M = \{(\boldsymbol{x}_s, y_s), s = 1, \ldots, S\}$ where $\boldsymbol{x}_s$ are drawn from the joint distribution of the inputs, $p(\boldsymbol{X})$, and $y_s = f(\boldsymbol{x}_s)$ are evaluations of the model output. We represent M as the matrix

$$M = \begin{pmatrix} x_{1,1} & x_{2,1} & \ldots & x_{i,1} & \ldots & x_{p,1} & y_1 \\ x_{1,2} & x_{2,2} & \ldots & x_{i,2} & \ldots & x_{p,2} & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1,S} & x_{2,S} & \ldots & x_{i,S} & \ldots & x_{p,S} & y_S \end{pmatrix}. \tag{29}$$

We then extract the x_i and y columns and then reorder this matrix row-wise with respect to x_i, giving

$$M_i^* = \begin{pmatrix} x_{i,(1)} & y_{(1)} \\ x_{i,(2)} & y_{(2)} \\ \vdots & \vdots \\ x_{i,(S)} & y_{(S)} \end{pmatrix}, \tag{30}$$

where $x_{i,(1)} \leq x_{i,(2)} \leq \ldots \leq x_{i,(S)}$, and $y_{(s)}$ is the model evaluated at $\boldsymbol{x}_{(s)}$.

Next, we divide the output $y_{(1)}, \ldots, y_{(S)}$ into K vectors, each of length b so that $S = Kb$, i.e. $\{y_{(1)}, \ldots, y_{(b)}\}, \{y_{(b+1)}, \ldots, y_{(2b)}\}, \ldots, \{y_{(S-b+1)}, \ldots, y_{(S)}\}$. The 'bandwidth' b is chosen to be small compared with the size of S.

We can obtain the main effect index either directly from the variance of the conditional expectation, or from the expectation of the conditional variance via the identity $\mathrm{var}_{X_i}\{E_{X_{-i}}(Y|X_i)\} = \mathrm{var}(Y) - E_{X_i}\{\mathrm{var}_{X_{-i}}(Y|X_i)\}$. Numerical stability of the algorithm with respect to the choice of d is improved if the expectation of the conditional variance, rather than the variance of the conditional expectation is approximated, and we therefore calculate $E_{X_i}\{\mathrm{var}_{X_{-i}}(Y|X_i)\}$, which we approximate as

$$E_{X_i}\{\mathrm{var}_{X_{-i}}(Y|X_i)\} \approx \frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{b}\sum_{j=b(k-1)+1}^{bk}\left(y_{(j)} - \bar{y}_k\right)^2\right\}, \tag{31}$$

where $\bar{y}_k = \frac{1}{b}\sum_{j=b(k-1)+1}^{dk} y_{(j)}$.

By re-ordering each M_i with respect to x_i, the main effect indexes for all p inputs can be obtained from a single Monte Carlo sample M.

Appendix C - Distribution for sum-to-one parameters

We denote a sum-to-one intermediate parameter as $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, where $Y_j \in [0,1]\ \forall j$ and $\sum_{j=1}^{n} Y_j = 1$.

The true unknown value of the intermediate parameter is denoted $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ where $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{\delta_Y}$ and $\boldsymbol{\delta_Y} = (\delta_{Y_1}, \ldots, \delta_{Y_n})$. The same constraints apply to $\boldsymbol{Z}$ as to $\boldsymbol{Y}$, i.e. $Z_j \in [0,1]\ \forall j$ and $\sum_{j=1}^{n} Z_j = 1$.

We state the following beliefs about $\boldsymbol{\delta_Y}$. Firstly, that $E(\delta_{Y_j}) = 0\ \forall j$, and secondly that the mean of the ratio of the standard deviation of the discrepancy to the expected value of the parameter is some constant v, i.e. that

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\sqrt{\text{var}(\delta_{Y_j})}}{E(Y_j)} = v. \tag{32}$$

We generate a sample from $p(\boldsymbol{Z})$ as follows. Firstly, we sample $\{\boldsymbol{y}_s, s = 1, \ldots, S\}$ from $p(\boldsymbol{Y})$. Conditional on $\boldsymbol{Y}$ we then generate a sample $\{\boldsymbol{z}_s, s = 1, \ldots, S\}$ from $p(\boldsymbol{Z})$, where each $\boldsymbol{z}_s$ is a single draw from $p(\boldsymbol{Z}|\boldsymbol{Y} = \boldsymbol{y}_s)$. The conditional distribution of $\boldsymbol{Z}|\boldsymbol{Y} = \boldsymbol{y}_s$ is Dirichlet with hyperparameter vector $\gamma\boldsymbol{y}_s = (\gamma y_{1,s}, \ldots, \gamma y_{n,s})$.

The expectation of δ_{Y_j} is

$$E(\delta_{Y_j}) = E(Z_j) - E(Y_j) = E_{Y_i}\{E_{Z_j}(Z_j|Y_j)\} - E(Y_j) = 0, \tag{33}$$

as required. The variance of δ_{Y_j} is

$$\begin{aligned}
\text{var}(\delta_{Y_j}) &= \text{var}(Z_j) + \text{var}(Y_j) - 2\text{Cov}(Z_j, Y_j) && (34)\\
&= E_{Y_j}\{\text{var}_{Z_j}(Z_j|Y_j)\} + \text{var}_{Y_j}\{E_{Z_j}(Z_j|Y_j)\} + \text{var}(Y_j) - 2\text{cov}(Z_j, Y_j) && (35)\\
&= E_{Y_j}\{\text{var}_{Z_j}(Z_j|Y_j)\} + \text{var}_{Y_j}\{E_{Z_j}(Z_j|Y_j)\} + \text{var}(Y_j) - 2\text{var}(Y_j) && (36)\\
&= E_{Y_j}\{\text{var}_{Z_j}(Z_j|Y_j)\} + \text{var}(Y_j) + \text{var}(Y_j) - 2\text{var}(Y_j) && (37)\\
&= E_{Y_j}\{\text{var}_{Z_j}(Z_j|Y_j)\} && (38)\\
&= E_{Y_j}\left\{\frac{(Y_j(1-Y_j)}{\gamma+1}\right\} && (39)\\
&= \frac{E(Y_j)\{1-E(Y_j)\}}{\gamma+1} - \frac{\text{var}(Y_j)}{\gamma+1} && (40)\\
&\approx \frac{E(Y_j)\{1-E(Y_j)\}}{\gamma+1}. && (41)
\end{aligned}$$

The final step follows because $\frac{\text{var}(Y_j)}{\gamma+1}$ is small relative to $\frac{E(Y_j)\{1-E(Y_j)\}}{\gamma+1}$.

The hyperparameter γ is chosen such that the mean of the ratio of the standard deviation to the expected value of the parameter is v, i.e. so that

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\sqrt{\text{var}(\delta_{Y_j})}}{E(Y_j)} = \frac{1}{n}\sum_{j=1}^{n}\frac{\sqrt{\frac{E(Y_j)\{1-E(Y_j)\}}{\gamma+1}}}{E(Y_j)} = v. \tag{42}$$

Approximating $E(Y_j)$ by the sample mean $\bar{y}_j$ and rearranging gives

$$\gamma = \frac{1}{v^2}\left\{\frac{1}{n}\sum_{j=1}^{n}\sqrt{\frac{1-\bar{y}_j}{\bar{y}_j}}\right\}^2 - 1. \tag{43}$$

References

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*, Chichester: John Wiley.

Bojke, L., Claxton, K., Sculpher, M. and Palmer, S. (2009). Characterizing structural uncertainty in decision analytic models: A review and application of methods, *Value Health*, **12 (5)**: 739–749.

Chilcott, J., Tappenden, P., Paisley, S., Kaltenthaler, E. and Johnson, M. (2010a). Choice and judgement in developing models for health technology assessment; a qualitative study, *ScHARR Discussion Paper*, Available from `http://www.shef.ac.uk/scharr/sections/heds/dps-2010.html`.

Chilcott, J., Tappenden, P., Rawdin, A., Johnson, M., Kaltenthaler, E., Paisley, S., Papaioannou, D. and Shippam, A. (2010b). Avoiding and identifying errors in health technology assessment models, *Health Technol Assess*, **14 (25)**.

Claxton, K., Sculpher, M., McCabe, C., Briggs, A., Akehurst, R., Buxton, M., Brazier, J. and O'Hagan, T. (2005). Probabilistic sensitivity analysis for nice technology assessment: not an optional extra, *Health Econ*, **14 (4)**: 339–347.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions, *J Am Stat Assoc*, **100 (470)**: 680–700.

Goldstein, M. and Rougier, J. (2009). Reified bayesian modelling and inference for physical systems, *J Stat Plan Inference*, **139 (3)**: 1221–1239.

Jackson, C. H., Sharples, L. D. and Thompson, S. G. (2010). Structural and parameter uncertainty in bayesian cost-effectiveness models, *J R Stat Soc Ser C Appl Stat*, **59 (2)**: 233–253.

Jackson, C. H., Thompson, S. G. and Sharples, L. D. (2009). Accounting for uncertainty in health economic decision models by using model averaging, *J R Stat Soc Ser A Stat Soc*, **172 (2)**: 383–404.

Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection, *J Am Stat Assoc*, **99 (465)**: 279–290.

Karnon, J. (2003). Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation, *Health Econ*, **12 (10)**: 837–848.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models, *J R Stat Soc Ser B Stat Methodol*, **63 (3)**: 425–464.

Kim, S.-Y., Goldie, S. J. and Salomon, J. A. (2010). Exploring model uncertainty in economic evaluation of health interventions: The example of rotavirus vaccination in vietnam, *Med Decis Making*, **30 (5)**: E1–E28.

Krahn, M. and Gafni, A. (1993). Discounting in the economic evaluation of health care interventions, *Medical Care*, **31 (5)**: 403–418.

NICE (2006). Four commonly used methods to increase physical activity: PH2, Tech. rep., NICE, Available from `http://www.nice.org.uk/PH2`.

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity of complex models: a Bayesian approach, *J R Stat Soc Ser B Stat Methodol*, **66**: 751–769.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006). *Uncertain Judgements: Eliciting Expert Probabilities*, Chichester: John Wiley and Sons.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*, Chichester: John Wiley and Sons Ltd.

Sonnenberg, F. A. and Beck, J. R. (1993). Markov models in medical decision making, *Med Decis Making*, **13 (4)**: 322–338.

Taylor-Robinson, D., Milton, B., Lloyd-Williams, F., O'Flaherty, M. and Capewell, S. (2008). Policy-makers' attitudes to decision support models for coronary heart disease: a qualitative study, *Journal of Health Services Research Policy*, **13 (4)**: 209–214.

Weinstein, M. C., O'Brien, B., Hornberger, J., Jackson, J., Johannesson, M., McCabe, C. and Luce, B. R. (2003). Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR task force on good research practices-modeling studies, *Value Health*, **6 (1)**: 9–17.