Bayesian Calibration of a Natural History Model with Application to a Population Model for Colorectal Cancer

Sophie Whyte, PhD, Cathal Walsh, PhD, Jim Chilcott, MSc

Background. *Cancer natural history models are essential when evaluating screening/preventative interventions or changes to diagnostic pathways. Natural history models commonly use a state transition structure, but it is often not possible to observe the state transition probabilities required for parameterization.* ***Aim***. *The work aimed to accurately represent the uncertainty in the parameters of a state transition model for the natural history of colorectal cancer by embedding the problem in the framework of Bayesian inference.* ***Methods***. *The Metropolis-Hastings algorithm was used to estimate natural history parameters and screening test characteristics by generating multiple sets of parameters from the posterior distribution, which is the probability distribution that is compatible with the observed data. Observed data included colorectal cancer incidence categorized by age and stage, autopsy data on polyp prevalence, and cancer and polyp detection rates from the first round of screening with the fecal occult blood test in England. The approach was implemented using Visual Basic.* ***Results***. *The results were subsequently examined for convergence using the package CODA in R 2.8.0. Outputs from fitting were samples from the joint posterior distribution of the natural history parameters given the epidemiological data. The parameter sets obtained are shown to have a good fit to all the observed data sets. These parameter sets are used when running probabilistic sensitivity analysis.* ***Conclusion***. *The advantages of this strategy are that it draws efficiently from a high-dimensional correlated parameter space. The algorithm is simple to code and runs overnight on a standard desktop PC. Using this method, the parameter sets are drawn according to their posterior probability given calibration data, and thus they correctly summarize the residual uncertainty in the parameter space.* ***Key words:*** *Bayesian; calibration; natural history; colorectal cancer.* **(*Med Decis Making XXXX;XX:xx–xx*)**

INTRODUCTION

Cancer natural history models are vital tools in developing our understanding of disease pathways and enabling improved decision making. These models commonly describe the disease natural history in terms of a set of mutually exclusive and exhaustive health states, together with a set of transition probabilities or rates that describe possible transitions between health states. There is currently a limited understanding of how best to estimate the set of transition probabilities within such natural history models. This article reports a Bayesian approach to model calibration that uses the Metropolis-Hastings algorithm to estimate transition probabilities and their joint uncertainty.

DOI: 10.1177/0272989X10384738

Cancer Natural History Models

Cancer natural history models are essential when evaluating screening interventions, changes to diagnostic pathways such as a new diagnostic tool, or preventative interventions. Such interventions aim to reduce morbidity, mortality, and treatment costs through earlier diagnosis and cancer prevention. Natural history models can predict the number of cases of cancer and precancerous conditions that are present in a population and the rate at which they are likely to progress. This information can then be used to estimate the capacity implications of an intervention (e.g., the number of positive screening

tests), for service planning, or to calculate the cost-effectiveness of a proposed intervention. In particular, modeling can be useful to help evaluate screening options when it is not feasible to perform large-scale, long-term trials of all options available.

Parameterizing a Natural History Model

Natural history models commonly use a state transition structure where the progression of disease is split into distinct health states. However, often the state transition probabilities required to parameterize such a model cannot be observed directly, and this can lead to difficulties. Specifically, in colorectal cancer (CRC), the rate at which a patient moves between the disease states would be extremely difficult to observe directly. There are very little data on the rate at which polyps occur and cancers grow as large polyps, and cancer will be treated if found for ethical reasons. It is known that polyps and early stage cancers are largely asymptomatic. The asymptomatic nature of early stage cancer is the motivation behind screening interventions. The probability that an existing cancer will be diagnosed cannot be directly measured, but there are data on the numbers who present for treatment and data from screening interventions. We therefore have related but not direct data from which we want to estimate a set of natural history parameters.

A model and a set of natural history parameters can be used to predict CRC incidence and screening outcomes. If model predictions are "close" to observed data, then we say that the parameter set has a "good fit." Pure optimization methods can be used to select a "best-fitting" set of natural history parameters; however, a method that explores the parameter space to identify plausible (i.e., reasonably well fitting but not necessarily the "best" fit) is needed for probabilistic sensitivity analysis.

A number of approaches have been taken by other authors working in this and related areas. For example, Tappenden and others[1,2] describe a multiple dimensional uniform simple Monte Carlo proposal set and accepted sets within a certain threshold on a goodness-of-fit test. Kim and others[3] use a likelihood-based approach to draw sets that are not "statistically significantly" different from the best-fitting set. Blower and others[4] used a Latin Hypercube design to sample the fit at different points and then used targets of fit to identify which combinations of parameters to use in further analyses, whereas Harper and Jones[5] used distributions guided by expert opinion. The Cancer Incidence and Surveillance Network (CISNET) group[6,7] described colorectal cancer models that use the Nelder and Mead simplex algorithm or an adaptation of this method to obtain an optimal parameter set. Finally, Rutter and others[8] have examined the statistical properties of methods such as this for similar calibration exercises in microsimulation models.

The existing Bayesian literature examines the problem of "black box" computationally intensive model calibration for many different examples. Some approaches to the problem, with examples, are provided by Higdon and others[9] and Goldstein and others.[10]

Bayesian Methods Applied to Cost-Effectiveness Analysis

A number of authors have examined the use of Bayesian methods in fitting models for cost-effectiveness analyses. This inferential framework is reported to be a convenient and natural way in which to combine evidence.[11] The residual uncertainty taking the evidence into account is encapsulated in the posterior distribution for the parameters. Samples from this joint distribution can then be used to quantify decision uncertainty, precisely as probabilistic sensitivity analysis (PSA) is used as advocated under the current National Institute for Health and Clinical Excellence (NICE) guidance.[12]

The software package WinBUGS can be used to implement the Bayesian approach using Markov chain Monte Carlo (MCMC). Examples include the evaluation of neuraminidase inhibitors for influenza, taxanes for the treatment of breast cancer, and

Received 21 July 2009 from the School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK (SW, JC); Department of Statistics, Trinity College Dublin, Dublin, Ireland (CW); and National Centre for Pharmacoeconomics, St. James's Hospital, Dublin, Ireland (CW). Thanks to Claire Nickerson of NHS cancer screening for providing data from the first round of screening in England. Thanks to Eva Morris, Matthew Day, and Ariadni Aravani at NYCRIS/NCIN for providing cancer incidence and survival data and to UKACR for providing cancer registry population denominators. The authors wish to acknowledge the whole team involved in the Irish HTA of CRC screening where this approach was initially developed. In particular, they would like to thank Linda Sharp, Alan O'Céilleachair, Lesley Tilson, and Cara Usher for input on the epidemiology and the identification of relevant literature and Paul Tappenden for advice on the existing model. The authors would like to thank the referees whose constructive comments were crucial to improving this article. Thanks also to Jason Madan, Mark Strong, and Matt Stevenson for commenting on early versions of this manuscript. Revised version accepted for publication 18 July 2010.

Address correspondence to Sophie Whyte, School of Health and Related Research (ScHARR), University of Sheffield, Regent Court, 30 Regent Street, Sheffield S14DA, UK; e-mail: Sophie.whyte@shef.ac.uk.

model calibration in its own right.[13,14] In addition to WinBUGS, it is possible to implement MCMC using other software, including Visual Basic within Excel, which is the approach taken here.

Aims and Objectives

Tappenden and others[1,15,16] used a Markov state transition model for the natural history of CRC in the option appraisal of population-based colorectal cancer screening programs in England that informed the establishment of the NHS Bowel Cancer Screening Programme. This model was updated and modified to evaluate possible screening options for Ireland.[17] This has subsequently been further developed and populated using recent English incidence data and data from the first round of screening in England, and it is this work that we describe here.[18] The recent availability of both incidence data with good staging information and results from screening in England makes this piece of work particularly timely.

The aim in updating the natural history model was to accurately represent the uncertainty in natural history parameters and fecal occult blood test (FOBT) characteristics. This was done by embedding the problem in the framework of Bayesian inference. Outputs from the fitting process were samples from the joint posterior distribution of the natural history parameters and FOBT characteristics given the epidemiological data. These parameter sets were used to run PSA for models evaluating screening programs, chemoprevention, and possible new screening/diagnostic technologies.

METHODS

Model Structure

The model uses a Markov state transition structure with a cycle length of 1 year and follows a cohort of 30-year-olds for 70 years. A diagram depicting the model structure is presented in Figure 1. The disease states are as follows:

- Normal epithelium
- Low-risk polyp
- High-risk polyp
- Dukes A CRC preclinical
- Dukes B CRC preclinical
- Dukes C CRC preclinical
- Stage D CRC preclinical
- Dukes A CRC clinical
- Dukes B CRC clinical

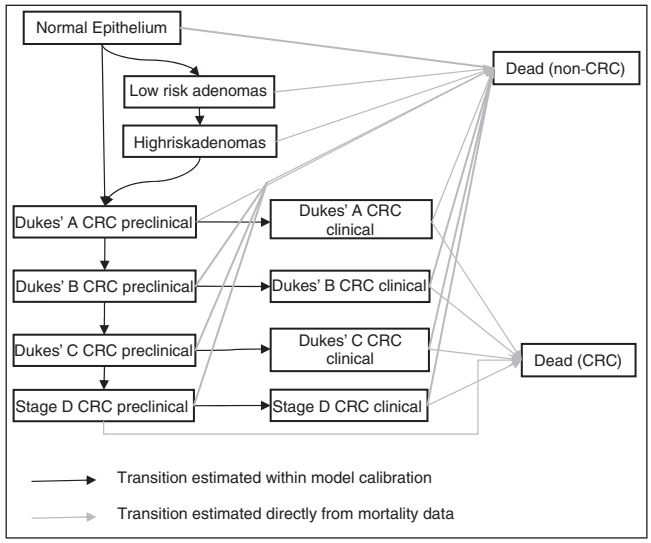

Figure 1 Model structure.

- Dukes C CRC clinical
- Stage D CRC clinical
- Dead

The low- and high-risk polyp health states are asymptomatic precancerous conditions, and if several polyps are present, then the most advanced will determine the health state.[19] CRC is split into 4 health states that describe the severity of the cancer. The preclinical CRC states are patients with undiagnosed CRC.

We define a sequence of annual transition probabilities between each of these states relating to CRC developing through the adenoma–carcinoma sequence, which is how most CRC is thought to arise.[20] In addition, we define a transition probability from normal epithelium to Dukes A CRC relating to cancers that do not arise from polyps, that is, de novo cancers. For each cancer state, we define a variable that is the probability of being diagnosed through symptomatic presentation or chance, and this corresponds to moving from a preclinical to a clinical health state. The model structure does not split cancers by proximal/distal colon and males/females. Including more detail in the modeling may make the structure more accurate, but the increased model complexity would necessitate additional unobservable input parameters.

All transition probabilities are assumed to be independent of age with the exception of the probability of developing a low-risk polyp. The probability of developing a low-risk polyp was allowed to vary by age. For simplicity, a piecewise linear model

Table 1 Values of Maximum a Posteriori Estimate and the Marginal 95% Credible Intervals for Each of the Natural History Parameters Estimated

Parameter	Maximum a Posteriori Estimate (95% Credible Interval Percentiles)
Normal epithelium to low-risk polyp, age 30	0.0019 (0.0003, 0.0036)
Normal epithelium to low-risk polyp, age 40	0.0010 (0.0004, 0.0026)
Normal epithelium to low-risk polyp, age 50	0.0053 (0.0030, 0.0068)
Normal epithelium to low-risk polyp, age 60	0.0110 (0.0071, 0.0157)
Normal epithelium to low-risk polyp, age 70	0.0156 (0.0091, 0.0153)
Normal epithelium to low-risk polyp, age 80	0.0019 (0.0011, 0.0126)
Normal epithelium to low-risk polyp, age 90	0.0046 (0.0017, 0.0143)
Normal epithelium to low-risk polyp, age 100	0.0074 (0.0008, 0.0129)
Low-risk polyp to high-risk polyp	0.1770 (0.1544, 0.1990)
High-risk polyp to Dukes A	0.0225 (0.0174, 0.0364)
Normal epithelium to Dukes A	3.04E-06 (1.60E-07, 3.4E-05)
Dukes A to Dukes B	0.91 (0.7306, 0.9696)
Dukes B to Dukes C	0.72 (0.7207, 0.9449)
Dukes C to Dukes D	0.83 (0.6073, 0.9194)
Probability of presenting symptomatically with Dukes A	0.09 (0.0694, 0.1006)
Probability of presenting symptomatically with Dukes B	0.21 (0.2148, 0.2693)
Probability of presenting symptomatically with Dukes C	0.49 (0.4115, 0.5269)
Probability of presenting symptomatically with stage D	0.82 (0.6109, 0.9891)
FOBT sensitivity for polyps	0.078 (0.0615, 0.1270)
FOBT sensitivity for CRC	0.354 (0.3278, 0.3923)
FOBT specificity	0.9945 (0.9940, 0.9952)

CRC, colorectal cancer; FOBt, fecal occult blood test.

was used with parameter values being the transition probabilities at ages 30, 40, . . ., 100.

All persons may die of causes other than CRC. Persons diagnosed with cancer may also die from CRC, and in addition, undiagnosed stage D may be fatal. Once a person is diagnosed with cancer, the transitions between Dukes stages are no longer modeled, and a stage-specific CRC mortality rate is applied. A significant proportion of patients survive colorectal cancer (5-year relative survival is over 90% for Dukes A), so it is not appropriate to use a constant mortality rate (exponential model). For each Dukes stage, a mixed model was used for CRC mortality, which assumes that a certain proportion of patients will be cancer survivors.

In addition to predicting transitions through the health states, the model is able to predict outcomes related to a screening intervention. The model uses FOBt sensitivity to polyps, FOBt sensitivity to CRC, and FOBt specificity to predict detection rates for both polyps and CRC in the prevalent round of screening in England.

Calibration Method

Natural history transition probabilities and FOBt characteristics are estimated within the calibration process, and these parameters are listed in Table 1. The Excel model could quickly be run to produce predictions of CRC incidence and screening outcomes for a given set of parameters. Our calibration process used the Metropolis-Hastings (MH) algorithm to generate multiple sets of parameters from the posterior distribution, that is, a probability distribution that is compatible with the observed data. The MH algorithm is conceptually straightforward and is described in Figure 2. An initial parameter set is selected at random. From this current parameter set, a proposal parameter set is obtained by making a random step in the parameter space. The likelihood of both the current set and the proposal set given the observed data is compared to determine whether the proposal parameter set is accepted. If the proposal parameter set is accepted, then it is the new current set. If the proposal set is not accepted, then the current parameter set is repeated in the chain. This process is then repeated to obtain a chain of parameter sets. During the initial "burn-in" period, the parameter sets in the chain will have decreasing likelihood, and the chain will converge to the part of the parameter space with the greatest likelihood. The MH random walk, which converges on the part of the parameter space with the greatest likelihood, is illustrated in the top section of Figure 3.

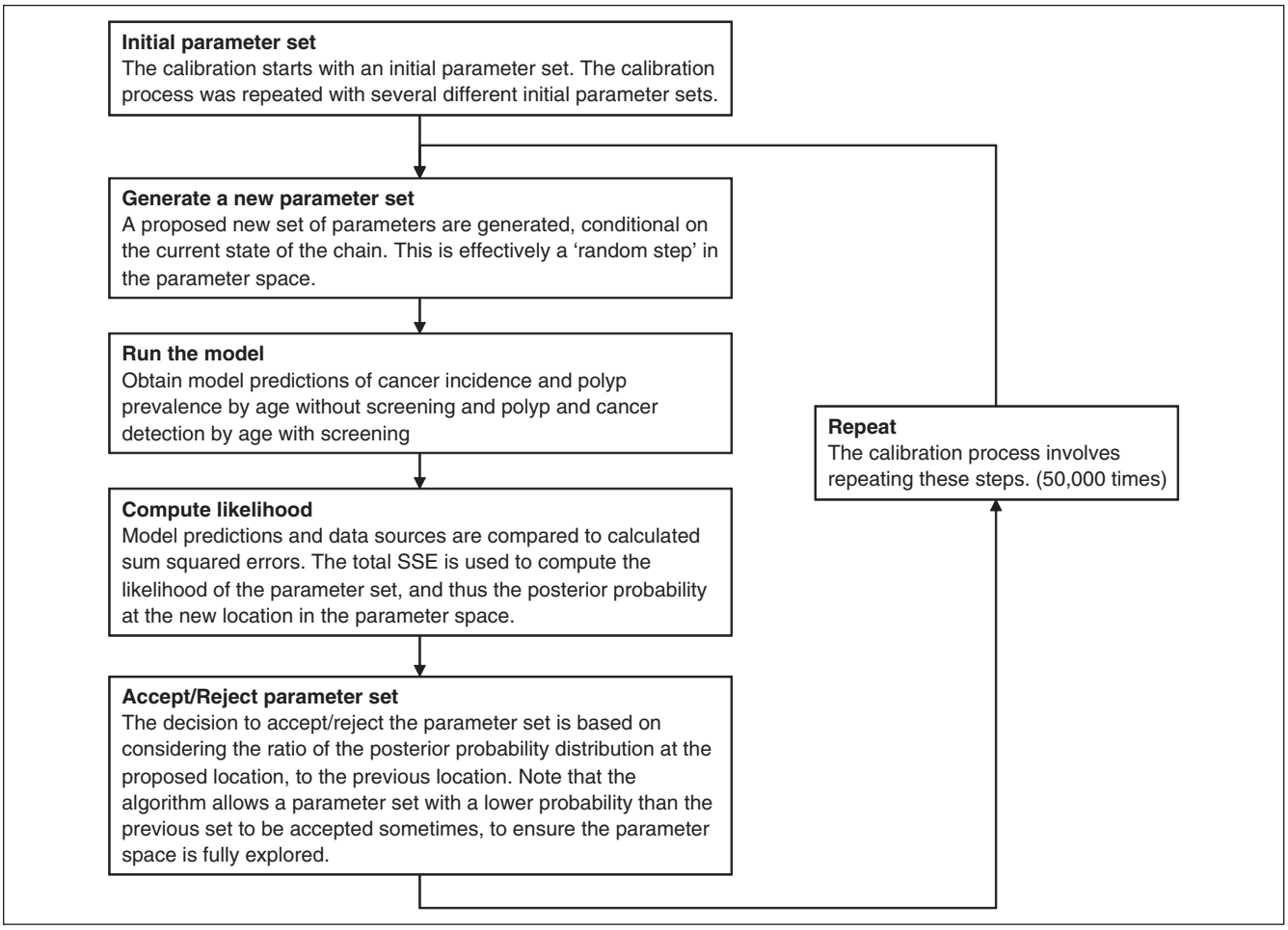

Figure 2 Flowchart describing the steps of the Metropolis-Hastings algorithm. SSE, sum of the squared errors.

The algorithm was easily coded in Visual Basic within Excel, and further details of the algorithm and implementation are provided in Appendixes A and B. As described by Higdon and others,[9] a direct MCMC approach may be taken in these circumstances and is a simple yet powerful method of drawing samples from the joint posterior distribution for the parameters. We will now describe in detail the steps and implementation of the MH algorithm and the calculation of the likelihood function.

The likelihood of the observed data, given model predictions, was taken to be normally distributed and centered on the fitted values. Thus, the joint posterior distribution factors as a product of Gaussians, and this is used during the acceptance step in the sampler.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{21})$ be a set of model parameters. Consider one observed data set, for example, Dukes A incidence. Let x_i denote the observed incidence at age i and μ_i denote the model-predicted incidence at that age, which is in turn a function of the model parameters, $\boldsymbol{\theta}$. We can consider the scatter of our observed incidence about the model as represented by independent Gaussians conditional on the parameters. The error term for these Gaussians is considered the same. Thus,

$$x_i \sim N(\mu_i, \sigma) \text{ for all ages, where } \mu_i = \mu_i(\boldsymbol{\theta})$$

and we denote the likelihood of observing x_i for a parameter set $\underline{\boldsymbol{\theta}}$ as

$$f(x_i \mid \boldsymbol{\theta}, \sigma) = f(x_i \mid \mu_i, \sigma).$$

Under the assumption that the errors are independent, conditional on the model, we can then factor this as follows:

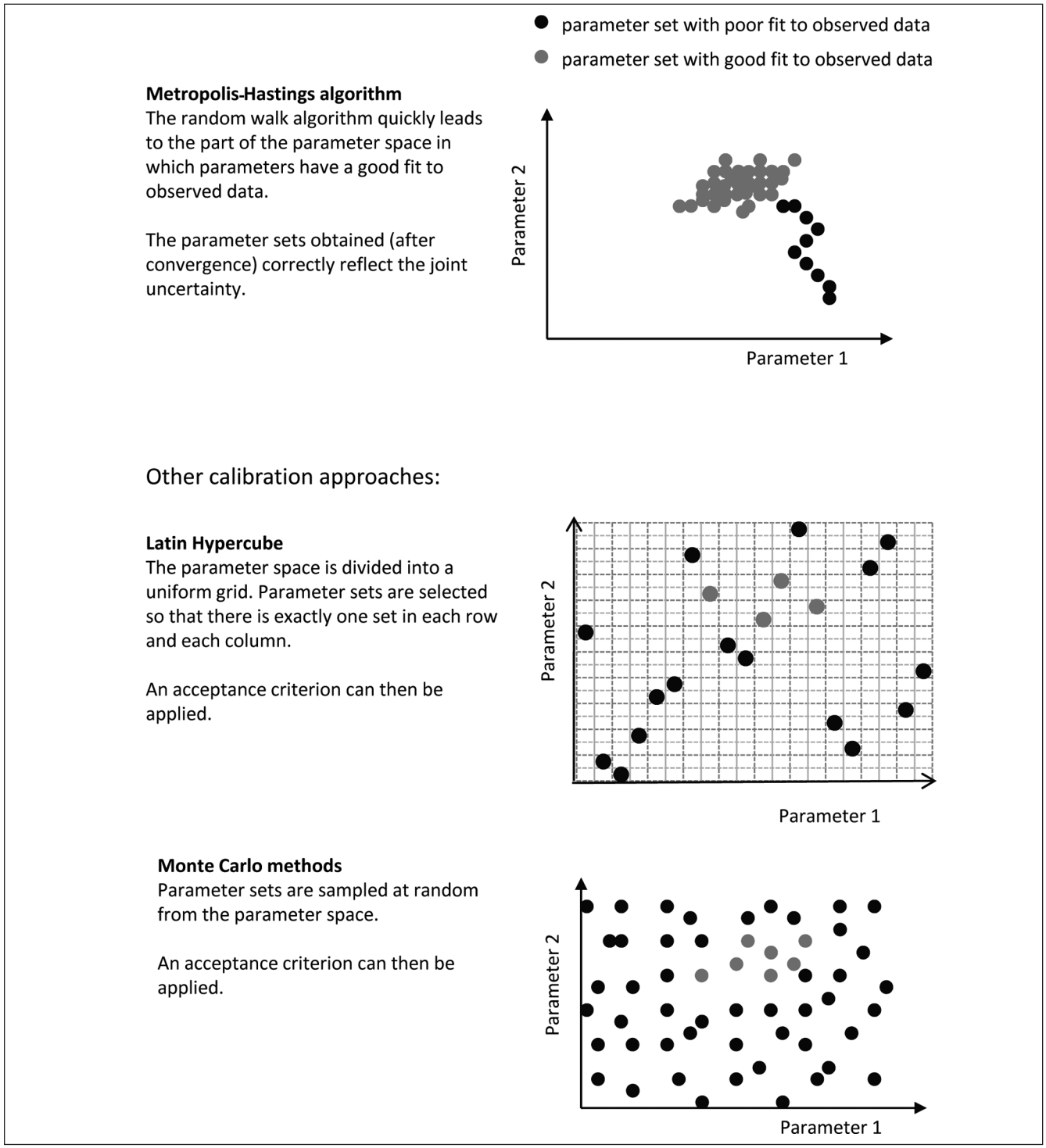

Figure 3 The Metropolis-Hastings algorithm and other calibration methods illustrated on a 2-dimensional parameter space.

$$f(x_1, x_2, x_3, \ldots, x_n \mid \mu_1, \mu_2, \ldots, \mu_n, \sigma) = f(x_1 \mid \mu_1, \sigma) f(x_2 \mid \mu_2, \sigma) \ldots f(x_n \mid \mu_n, \sigma).$$

Thus, this can be written using the probability density function of a Gaussian as

$$f(\mathbf{x} \mid \boldsymbol{\theta}, \sigma) = (2\pi\sigma^2)^{(n/2)}\exp(-(2\sigma^2)^{-1}((x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 + \ldots + (x_n - \mu_n)^2)),$$

where the term

$$(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 + \ldots + (x_n - \mu_n)^2$$

is the sum of the squared errors (SSE) of the difference between the observed data and prediction from the model. So the likelihood function is proportional to

$$\exp(-0.5(\mathrm{SSE}/\sigma^2)),$$

and this formula is used in the acceptance step of the algorithm.

The calibration fitted to 9 observed data sets: adenoma prevalence, screening number of positive FOBt, screening number of low-risk and high-risk (HR) polyps detected at colonoscopy, screening number of CRC detected at colonoscopy, Dukes A incidence, Dukes B incidence, Dukes C incidence, and stage D incidence. For each data set, the SSE is calculated be comparing the observed number of observations to the predicted number of observations for each age.

The target distribution for our analysis is the posterior distribution for the parameters, and this requires specification of a prior distribution as well as the likelihood function. In general, if there is some information about the parameters from other sources, this can be used directly in the formulation of the prior, but in this calibration, noninformative Beta(1,1) priors were used for all parameters. The proposal set is obtained by varying the values of each parameter in the current set by a small amount. A different increment, epsilon, is used for each parameter. For each parameter in the set, the new proposal value is obtained by adding a random sample from Uniform $(-\varepsilon_i, \varepsilon_i)$ distribution to the current parameter value. The proposal distributions for the sampler were tailored so that the acceptance rate for the algorithm was approximately 0.23.[21] This was done by examining initial outputs from the chains for each parameter and then increasing the width of the proposal, ε_i, where the acceptance rate was high and decreasing it where the acceptance rate was low.

To run the calibration, an estimate of variance is required for each observed data set, and these were obtained as follows. For each data set, an estimate of the minimum SSE value was obtained by running the MH algorithm 1000 times using only the SSE relating to that data set to determine acceptance. For each data set, the minimum SSE value was divided by the number of observations in the data set to provide an estimate of variance. To aid quick convergence, the values for the variances were held constant during the calibration process. The variance parameters for the Gaussians associated with the observed data sets enter the joint likelihood function as weights for each of the fitted curves. The total SSE was calculated as

$$\text{Total SSE} = \sum_{j=1..9} \frac{\text{SSE data set j}}{\text{variance data set j}}.$$

The acceptance step of the MH algorithm compares the total SSE between the current and proposal parameter sets. If the proposal parameter set has lower total SSE than the current parameter set, then it will be accepted. Otherwise, the proposal set will be accepted with probability:

$$\exp(-0.5(\text{Total SSE}_{\text{proposal set}} - \text{Total SSE}_{\text{current set}})).$$

We note that in this calibration, we have weighted all data sets equally (but the algorithm implicitly weights according to sample size), but if there were significant differences between the quality of the data in different sets, then it may be appropriate to weight data sets accordingly.

The MH algorithm was run using 6 different initial parameter sets, and these runs are referred to as independent chains. Chains were run for 50,000 iterations with a burn-in of 10,000 iterations and were thinned so that every fifth parameter set was stored. The model calibration process was undertaken using standard desktop PCs (2.66-GHz Core 2 Duo processor, 3 GB RAM), and the code was run overnight, which was facilitated by having free access to multiple machines in an undergraduate teaching laboratory. The results were examined for convergence using the package CODA in R 2.8.0.[22] Autocorrelation and cross-correlation for the chains were also examined using the standard diagnostic suite within CODA.

Parameter sets from after convergence in each of the 6 chains were combined, and this combined set was used to calculate the a posteriori estimate and the 95% credible interval for each parameter. A sample of these parameter sets was used for probabilistic sensitivity analysis. Since the values are samples from the joint posterior probability distribution, they reflect the residual uncertainty about the natural

history parameters conditional on the data that are available for the fitting process.

Data Sources

The model was calibrated to incidence and polyp prevalence data in the absence of screening and also to results from the prevalent round of the NHS Bowel Cancer Screening Programme in England. As the cohort of individuals in the model ages from 30 to 100, the model predictions are compared to observed data at each age.

England has the advantage of having a particularly good data set for such a calibration. From 2004 to 2006, there was virtually no screening, and there are incidence data that include stage at diagnosis. After 2006, population screening commenced, and there is now a large data set from the first round of screening.

Incidence data were grouped by both age at diagnosis (quinary age bands) and stage at diagnosis. It is important that data split by age and stage are used, as this will inform the transition rates between stages. Data from Oxford, Northern and Yorkshire, and Eastern regions from 2004 to 2006 was used. These regions were chosen as they have a good level of recording of staging information, with a proportion unstaged of 22% when combined. Incidence data from after 2006 cannot be used, as screening was taking place then, so the underlying natural history is affected. Unstaged cases were assumed to be Dukes C or stage D, and a 55%:45% split was applied as this is the split in the staged cases. This assumption is supported by evidence that survival in the unstaged cases is between that of C and D. Incidence data were provided by the Northern and Yorkshire Cancer Registry and Information Service (NYCRIS) on behalf of the National Cancer Intelligence Network (NCIN).[23] The figures were based on data that were collected and quality assured by the 8 regional English cancer registries. The Cancer Registry population denominators were provided by the United Kingdom Association of Cancer Registries (UKACR).[24]

The proportion of total CRC incidence that arises in persons with irritable bowel disease (IBD) is 2%, and expert opinion suggests that the same proportion of CRC may arise from other non-polypoid routes.[25] Based on this information, the proportion of cancers that arise from non-polypoid routes, which we refer to as de novo CRC, was represented by a prior beta distribution with a mean of 5% and a wide 95% confidence interval (1%–13%).

As colonoscopy is an invasive procedure associated with significant morbidity, there are limited data available on underlying polyp prevalence. Pendergrass and others[26] report on a well-conducted autopsy study from the Mayo clinic (N = 2807) describing prevalence of adenomatous polyps in 30- to 89-year-olds split into 10-year age bands. In addition, data from several smaller and older autopsy studies were included within the calibration.[27–32] A high level of uncertainty was present in the polyp prevalence data due to the age of the studies, geographical differences in populations, and weaknesses in study designs.

Detection rates for cancer, high-risk and low-risk polyps, and FOBt positivity rates in the first (prevalent) round of screening in England categorized by ages 60 to 69 were used within the model calibration.[18] This data set includes 1.9 million persons who have attended CRC screening in England.

Model predictions for screening outcomes are highly sensitive to screening test characteristics, and there is considerable uncertainty around FOBt characteristic values. For this reason, the FOBt characteristic values were estimated within the calibration process.

Mortality rates were not estimated within the calibration process; instead, they were estimated directly from available data. Age-specific other-cause death was estimated by removing CRC deaths from all-cause mortality taken from Office of National Statistics life tables.[33] CRC survival rates dependent on age, CRC stage at diagnosis, and time since diagnosis were used. Data consisted of CRC 1-, 3-, and 5-year relative survival from persons diagnosed from 1997–2001 in England (provided by NYCRIS and the NCIN).[23] CRC survival was estimated to be relative survival multiplied by the age-specific expected survival.

The data on 1-, 3-, and 5-year survival were extrapolated using a mixed model in which a proportion of patients are assumed to have terminal cancer with exponential survival, whereas remaining patients are assumed to be cancer survivors. The proportion of patients who are cancer survivors was estimated to be slightly lower than the 5-year survival rates. This is based on clinical opinion, which suggests that for patients who have survived CRC for 5 years, the mortality rate for subsequent years is very low.

RESULTS

The output of the chains was saved in Excel and then exported to .csv format. These files were then

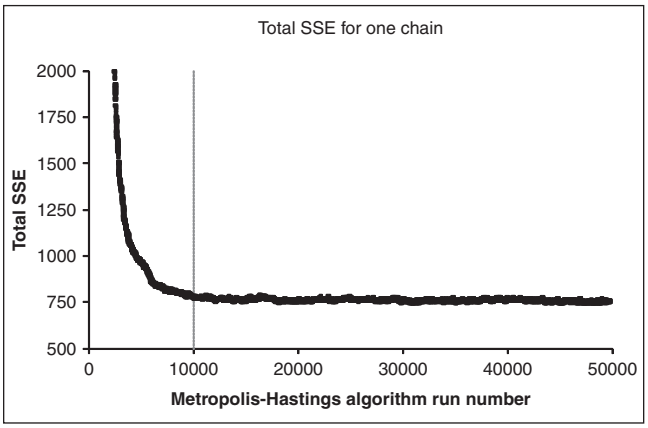

Figure 4 The total sum of the squared errors (SSE) during 1 run of the calibration process.

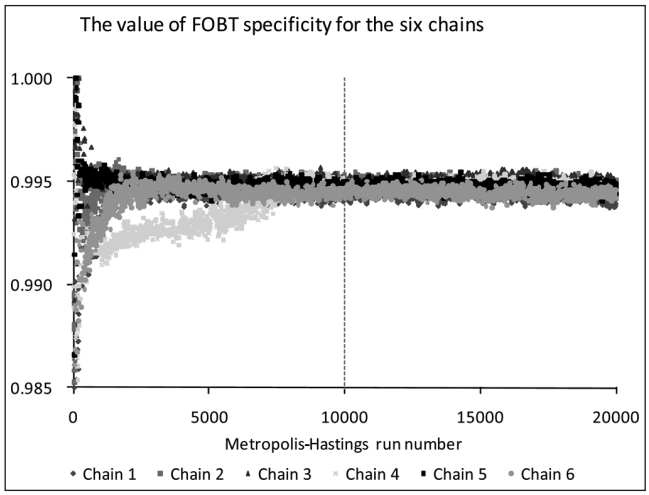

Figure 5 How 6 chains explore the parameter space for 1 parameter. FOBT, fecal occult blood test.

read into R and exported as text in list format, which is the same as produced by BUGS. The index file was also created for the variables. The chains were then processed using codamenu() in CODA v.0.13-4.[22] The Gelman and Rubin plots for each of the variables were examined. Formal diagnostics examined included the interval estimates of potential scale reduction factors, Raftery and Lewis convergence diagnostics, and the Heidelberger and Welch stationarity and interval half-width tests. Graphical summaries of the chains were also examined, and examples of these are shown in Figures 4 and 5.

Figure 4 shows the total SSE (as defined earlier) for one of the chains. At the start of the chain, as convergence occurs, the total SSE is quickly reduced. In the later part of the chain, the total SSE may increase as the algorithm can allow a step to a parameter set with a higher total SSE. Figure 5 shows the 6 chains for the FOBT specificity. Each of the 4 chains is exploring the parameter space. Convergence was assessed using CODA in R as described earlier.

Table 1 shows the maximum a posteriori estimate of the parameter values, that is, the best-fitting parameter set. Also presented are the 95% credibility intervals for each parameter, which were calculated using parameter sets after burn-in.

Figures 6, 7, and 8 compare model predictions to observed data for CRC incidence, polyp prevalence, and screening outcomes. Rather than plotting the model predictions corresponding to just one parameter set, we present the range of model predictions that corresponds to the parameter sets obtained in the calibration process. The range of model predictions based on 120 values randomly selected from the posterior predictive are presented. This corresponds to an approximate 99% prediction interval. Using these parameter sets represents the uncertainty in the model predictions. The graphs demonstrate that overall, the results of the calibration had a good fit to each of the observed data sets.

Figure 6 shows CRC incidence rate by age and stage. The uncertainty in the predictions is greater in ages older than 75 due to a smaller population and an absence of screening data for these ages.

Figure 7 shows polyp prevalence by age and demonstrates that there is considerable variation between the prevalence seen in the different autopsy studies. The figure shows that there is considerable uncertainty in model-predicted polyp prevalence. The predicted prevalence values are roughly between the values from the 2 largest autopsy studies. It is important to realize that, as the screening data have large sample sizes, the polyp detection rates seen at screening will have a bigger impact on the predicted polyp prevalence rates than the smaller autopsy studies. Due to the heterogeneity, the parameter estimates may be sensitive to inclusion of particular studies; however, this was not explored in depth due to the computational overhead.

Figure 8 shows the model predictions for the 4 screening outcomes: CRC and low/high-risk polyp detection rates and FOBT positivity rates. A very good fit is seen in all ages except for 70 to 74. As the

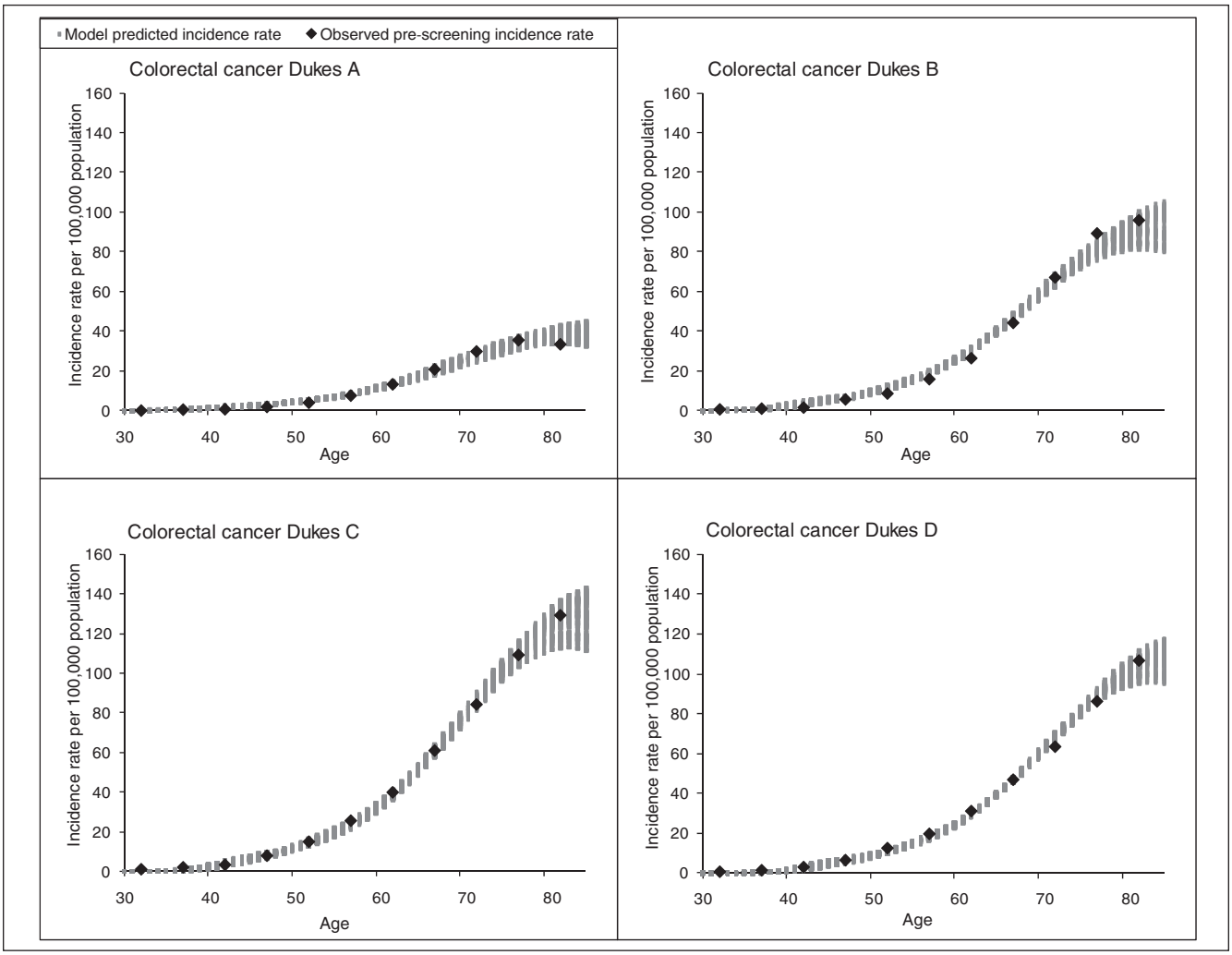

Figure 6 Colorectal cancer (CRC) incidence rate in the absence of screening, by age and Dukes stage at diagnosis: model predictions compared to observed data.

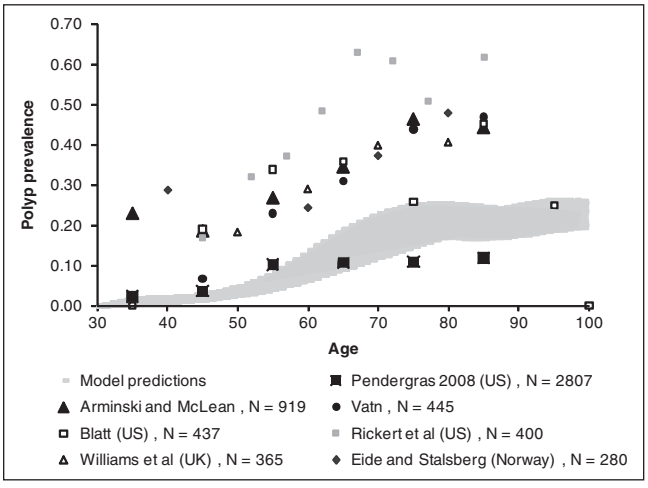

Figure 7 Polyp prevalence: data from autopsy studies and model predictions.

number of observations in the 70 to 74 age group was very low, the rates seen in this age group are subject to significant sampling uncertainty.

Figure 9 shows a cross-plot of 2 parameters: the transition probability from low-risk to high-risk polyp and the transition probability from high-risk polyp to Dukes A preclinical. The figure shows that there is a positive correlation between these 2 parameter values. The correlation matrix demonstrates that there is correlation between many of the parameters (see Appendix C). The correlation between parameters confirms the importance of sampling natural history parameters from the joint distribution. This is a good example of why the independent (uncorrelated) distributions alone are an inappropriate summary for natural history parameters such as these. The importance of using correlated parameter sets can be demonstrated by

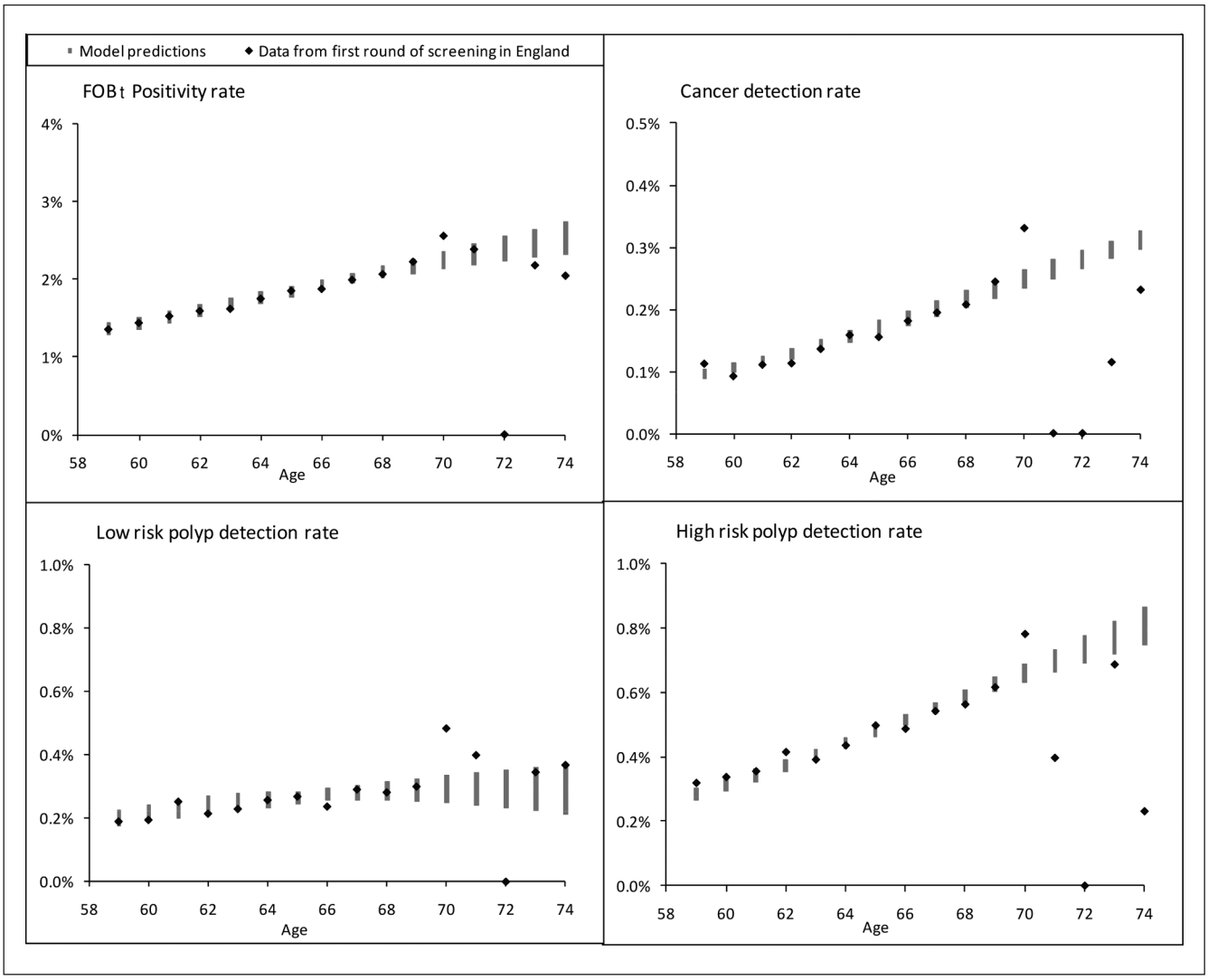

Figure 8 Screening outcomes from the prevalent round: data compared with model predictions. FOBT, fetal occult blood test.

sampling parameters independently from the marginal distributions. Figure 10 shows model predictions for 120 parameter sets (20 sets from each of the 6 runs) independently sampled from the marginal distributions. It is clear that some of the sampled parameter sets have a poor fit to the observed data.

There is considerable uncertainty surrounding the test characteristics of FOBT, which was the motivation for including these parameters within the calibration.[34] We observe that the estimates of FOBt sensitivity produced by this calibration seem consistent with values seen in trial situations. A systematic review and pooled analysis of the sensitivity of guaiac FOBT (gFOBT) reported a 95% confidence interval of 0.31 to 0.42 for cancer and 0.10 to 0.12 for polyps, whereas the calibration produced a 95% credible interval of 0.33 to 0.39 for cancer and 0.06 to 0.13 for polyps.[17] This alignment between sensitivity in model predictions and trial data provides a validation of the model.

The specificity credible interval obtained from the calibration of 0.994 to 0.995 is higher than the range 0.96 to 0.98 obtained in the systematic review. This difference could be caused by differences between the England screening population and the trial populations or by differences in the use of the FOBT test such as the referral algorithm.[35]

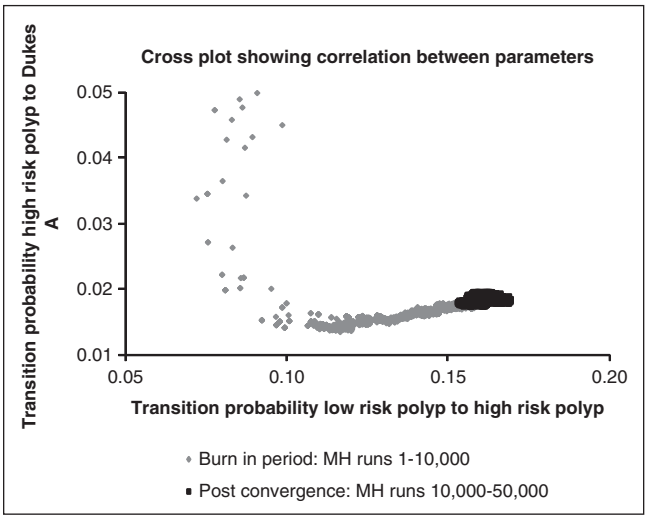

Figure 9 Cross-plot of 2 of the natural history parameters. The starting values for the chain are at the top left of the plot. The burn-in period of the chain is represented in gray by the trail from the top left to the main body of the distribution. The diagonal cigar shape shows strong correlation between these parameters and emphasizes the importance of treating these parameters as correlated "sets." MH, Metropolis-Hastings algorithm.

CONCLUSIONS AND DISCUSSION

Comparison with Other Methods

A comparison with the results of the previous calibration method is presented in Appendix D. The previous calibration method used a multiple dimensional uniform simple Monte Carlo proposal set and accepted sets within a certain threshold on a goodness-of-fit test.[2,36] Comparison with the results of the previous calibration is of limited value, as different data were used to calibrate the model, and changes have been made to the model structure. The previous work did not include age-dependent probabilities for the transition from normal epithelium to low-risk polyp, did not allow for de novo cancers, and did not estimate test characteristics within the calibration process. Some of the parameter estimates differ significantly between the methods. We suggest that the difference in FOBT characteristic values could be the cause of the differing results.

In comparison with the Latin Hypercube approach, as used by Blower and others,[4] and the approach used by Tappenden and others,[1,2] the MH algorithm is a more efficient way of exploring a high-dimensional parameter space with sets being sampled according to how well they fit the observed data.

Kim and others[3] used a likelihood-based approach to draw sets that are not statistically significantly different from the best-fitting set. By ensuring that the sets that are chosen do not differ from the best-fitting set, the correlation between parameters will be taken into account. This approach is a 2-step approach (generating sets and then observing fit), which compares to the approach taken here, which generates correlated sets of parameters directly within the MCMC.

Previous approaches have frequently used distributions guided by expert opinion, for example, Harper and Jones.[5] As data on incidence and screening results are now available, it is not necessary to resort to parameter estimates informed solely by expert opinion. However, within a Bayesian framework as used here, expert opinions can be used to elicit priors for the parameters.

A Bayesian approach using the MH algorithm was used in the microsimulation model by Rutter and others.[6,8] A cohort model approach is less computationally intensive compared with microsimulation.

Figure 3 illustrates the difference between some of these approaches for a 2-parameter space. This figure demonstrates that the MH algorithm is "more efficient" in the sense that it directly gets at the values rather than having to simulate lots of proposals at a first step then filter at a second. As for disadvantages, it does require careful consideration of convergence, and experience in doing this is useful/important in practice.

Discussion of Model Structure

A more complex CRC natural history model structure could be used: for example, transition probabilities could be varied by birth cohort, sex, or adenoma location. It is possible to include many more parameters in the model and estimate them as described. However, as with all modeling, there is a tradeoff between model complexity and the data available to identify values for these parameters. The parameter set was chosen with this tradeoff in mind and by considering the data that were available at the time and from studies under way. Although clinical opinion was important in developing the model structure, it is worth noting that no clinical opinion was required to parameterize the model. As more detailed data become available, refinement of the model structure is anticipated for future applications.

The transition probabilities from normal to low-risk polyp for each age were assumed to be

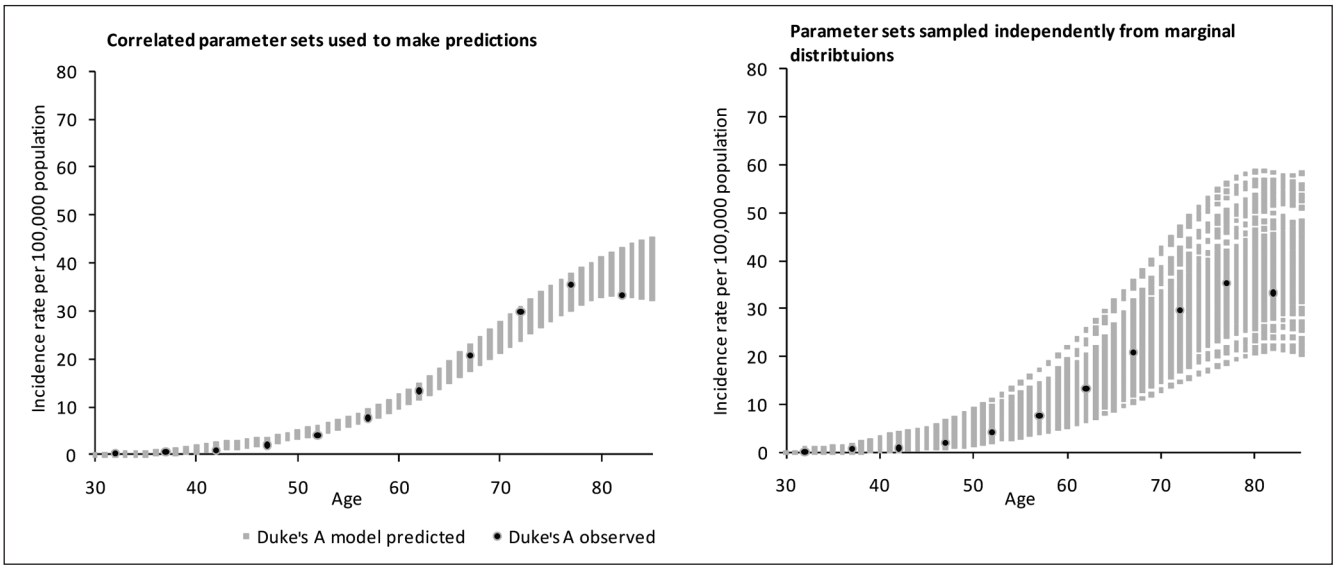

Figure 10 Dukes stage A colorectal cancer (CRC) incidence rate in the absence of screening, by age and Dukes stage at diagnosis: model predicted compared to observed data.

independent, although another approach could have constrained these in some way. This independence allowed the model to reflect the heterogeneity within the cohort. These values represent the average probability of transition from normal to low-risk polyp for the cohort; however, the cohort contains individuals with varying risks of polyp growth. It may be the case that after a certain age, the majority of high-risk individuals will have grown polyps—hence the population remaining in the "normal epithelium" disease state will consist of individuals with a lower risk of polyp growth. The data on polyp prevalence from the study by Pendergrass and others[26] seem to plateau in older ages, and this fits with the suggestion that the probabilities of transition from normal to low-risk polyp do not increase monotonically with age.

The annual transition probabilities used within this model may be a limitation. As one of the estimated transition probabilities is 0.91, this suggests that a more accurate representation may be obtained if a shorter cycle length is used.

Advantages of Approach

Using this method, the parameter sets are drawn according to their posterior probability given calibration data, and thus they correctly summarize the residual uncertainty in the parameter space given the data that are available. The strong correlation seen between parameters confirms the importance of this strategy, which ensures that candidate sets that are sampled conform correctly to this correlation structure. The correlation seen demonstrates the importance of sampling natural history parameters from the joint distribution when running PSA to consider uncertainty.

The underlying methodology is well developed in the Bayesian literature. The algorithm is simple to code in most programming languages (in this case, Visual Basic). Open-source modules for performing diagnostics on the outputs are freely available (CODA). We acknowledge that some prior experience of MCMC is useful to run the calibration effectively; however, this experience requirement holds for most computational methods used in calibration experiments.

APPENDIX A
Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a Markov chain Monte Carlo method, developed in the early computational science literature, which has been used extensively in Bayesian computation for many years. A simple explanatory article in the statistics literature is developed by Chib and Greenberg.[37]

The algorithm sets about constructing a Markov transition kernel that has as its equilibrium distribution some target density (in this case, the posterior distribution for the natural history parameters) of interest to the operator.

Using the notation used earlier in this article, the set of model parameters of interest is denoted by the vector, $\boldsymbol{\theta}$, and the objective is to sample from the posterior distribution of these parameters, given the data of observation, $\mathbf{x}$. At each stage of the sampler, a vector of parameters $\mathbf{z}[i]$ is recorded. After convergence, the algorithm provides that $\mathbf{z}[j] \sim f(\boldsymbol{\theta} \mid \mathbf{x})$.

The proposal distribution, q(), moves from a given state vector to a nearby location in the joint parameter space. Here the proposal is multivariate uniform with tuned variance.

The algorithm, then, works as follows:

Set i = 0; Set N = 500000;
Choose an initial state for $\mathbf{z}[0]$.
Propose $\mathbf{y}$ from $q(\mathbf{y} \mid \mathbf{z}[i])$.
Accept the proposal with probability
$= \min(1,(f(\mathbf{y} \mid \mathbf{x})q(\mathbf{z}[i] \mid \mathbf{y}) / f(\mathbf{z}[i] \mid \mathbf{x})q(\mathbf{y} \mid \mathbf{z}[i])))$.
If accepted, set $\mathbf{z}[i+1]=\mathbf{y}$, else set $\mathbf{z}[i+1]=\mathbf{z}[i]$.
If i<N set i=i+1; and go back to step 2.

After running the algorithm, it is essential to run diagnostic tests on the output to ensure that the algorithm has converged and that the autocorrelation of the chains is correctly taken into account when using the outputs. A review of standard diagnostics is available in Cowles and Carlin[38] and Brooks and Roberts.[39] A package in R, CODA,[22] is available that facilitates the running of these diagnostics in a straightforward fashion.

APPENDIX B
VBA Code Used to Implement Metropolis-Hastings Algorithm within Excel

```
Sub MetropolisHastings()

'Enters initial values as the current set.
Calculate
Range("initialset1").Select
Selection.Copy
Range("currentset").Select
Selection.PasteSpecial Paste:=xlValues

'Says how many iterations to run.
Do While Range("Current") < Range("max") + 1
Application.StatusBar = "run no. " & Range("current")


'Takes samples from the proposal distribution (named
cells on worksheet)
Range("ProposalSetA").Select
Selection.Copy
Range("ProposalSet").Select
Selection.PasteSpecial Paste:=xlValues

'Recalculate the model with new parameter set to yield
new prediction set
Calculate

'Decide whether to accept or reject the new parameter
set. The acceptreject value is obtained by comparing
the acceptance probability to a rand() according to the
MH algorithm as described in Appendix 1

If Range("acceptreject").Value = 1 Then
Range("PROPSETentire").Select
Selection.Copy
Range("CurrentSET").Select
Selection.PasteSpecial Paste:=xlValues
End If

'Pastes the current parameter set onto the list of param-
eter sets. This stores the values z[i] on the worksheet

Range("currentset").Select
Selection.Copy
Range("PasteCaseHere").Select
ActiveCell.Offset((Range("current"), 0).Select
Selection.PasteSpecial Paste:=xlValues
End If

Range("current") = Range("current") + 1
Loop
End Sub
```

APPENDIX C
Correlation Matrix

Correlation Matrix	Normal epithelium to low-risk polyp, age 30	Normal epithelium to low-risk polyp, age 40	Normal epithelium to low-risk polyp, age 50	Normal epithelium to low-risk polyp, age 60	Normal epithelium to low-risk polyp, age 70	Normal epithelium to low-risk polyp, age 80	Normal epithelium to low-risk polyp, age 90	Normal epithelium to low-risk polyp, age 100	Low-risk polyp to high-risk polyp	High-risk polyp to Dukes A	Normal epithelium to Dukes A	Dukes A to Dukes B	Dukes B to Dukes C	Dukes C to Dukes D	Probability of presenting symptomatically with Dukes A	Probability of presenting symptomatically with Dukes B	Probability of presenting symptomatically with Dukes C	Probability of presenting symptomatically with Dukes D	FOBT sensitivity for polyps	FOBT sensitivity for CRC	FOBT specificity
Normal epithelium to low-risk polyp, age 30	1																				
Normal epithelium to low-risk polyp, age 40	−0.639	1																			
Normal epithelium to low-risk polyp, age 50	0.841	−0.430	1																		
Normal epithelium to low-risk polyp, age 60	0.791	−0.108	0.860	1																	
Normal epithelium to low-risk polyp, age 70	0.021	0.382	0.167	0.255	1																
Normal epithelium to low-risk polyp, age 80	−0.570	0.042	−0.665	−0.743	−0.301	1															
Normal epithelium to low-risk polyp, age 90	−0.724	0.138	−0.790	−0.854	−0.411	0.787	1														
Normal epithelium to low-risk polyp, age 100	−0.012	−0.372	−0.274	−0.341	−0.469	0.171	0.477	1													
Low-risk polyp to high-risk polyp	−0.738	0.173	−0.882	−0.879	−0.103	0.705	0.773	0.362	1												
High-risk polyp to Dukes A	−0.673	−0.006	−0.803	−0.905	−0.391	0.735	0.846	0.542	0.889	1											
Normal epithelium to Dukes A	−0.291	0.034	−0.095	−0.201	−0.136	0.093	0.227	−0.027	0.134	0.193	1										
Dukes A to Dukes B	−0.072	0.286	0.139	0.172	0.431	−0.318	−0.297	−0.579	−0.201	−0.396	0.089	1									
Dukes B to Dukes C	0.149	−0.235	0.061	0.057	−0.236	0.295	0.070	−0.033	−0.010	0.002	0.104	0.281	1								
Dukes C to Dukes D	−0.310	0.416	−0.075	−0.049	0.375	−0.339	−0.068	−0.370	−0.033	−0.205	0.117	0.849	−0.090	1							
Probability of presenting symptomatically with Dukes A	−0.010	0.014	−0.067	−0.044	−0.304	−0.134	0.006	−0.081	−0.220	−0.237	−0.048	0.155	−0.070	0.258	1						
Probability of presenting symptomatically with Dukes B	0.098	−0.144	0.025	0.022	−0.484	0.076	0.059	−0.048	−0.228	−0.167	0.039	0.213	0.550	0.062	0.731	1					
Probability of presenting symptomatically with Dukes C	−0.306	0.386	−0.137	−0.102	0.285	−0.309	−0.071	−0.371	−0.045	−0.213	0.061	0.784	−0.124	0.930	0.521	0.260	1				
Probability of presenting symptomatically with Dukes D	0.087	0.017	0.069	0.089	0.133	0.245	−0.059	−0.101	0.007	0.026	−0.103	−0.676	−0.360	−0.708	−0.334	−0.435	−0.679	1			
FOBT sensitivity for polyps	−0.694	−0.007	−0.848	−0.948	−0.477	0.750	0.888	0.528	0.866	0.961	0.207	−0.374	0.011	−0.161	0.022	0.028	−0.103	−0.042	1		
FOBT sensitivity for CRC	0.211	0.046	0.390	0.397	0.216	−0.146	−0.302	−0.539	−0.358	−0.461	0.085	0.729	0.658	0.402	−0.099	0.299	0.263	−0.406	−0.479	1	
FOBR specificity	−0.484	−0.003	−0.509	−0.630	−0.193	0.572	0.593	0.310	0.614	0.722	0.172	−0.242	0.044	−0.143	−0.372	−0.259	−0.217	0.069	0.653	−0.208	1

CRC, colorectal cancer; FOBT, fecal occult blood test.

APPENDIX D
Comparison with Results of a Previous Calibration

Parameter	Previous Calibration (Tappenden et al.[15])	Results of This Calibration: Maximum a Posteriori Estimate (95% Credible Interval)
Normal epithelium to low-risk polyp	0.016	0.0010-0.0156 (age dependent)
Normal epithelium to Dukes A	0 (assumption)	3.04E-06 (1.60E-07, 3.40E-05)
Low-risk polyp to high-risk polyp	0.021	0.1770 (0.1544, 0.1990)
High-risk polyp to Dukes A	0.033	0.0225 (0.0174, 0.0364)
Dukes A to Dukes B	0.583	0.91 (0.73, 0.97)
Dukes B to Dukes C	0.656	0.72 (0.72, 0.94)
Dukes C to stage D	0.865	0.83 (0.61, 0.92)
Presenting symptomatically with Dukes A	0.07	0.09 (0.069, 0.101)
Presenting symptomatically with Dukes B	0.32	0.21 (0.215, 0.269)
Presenting symptomatically with Dukes C	0.49	0.49 (0.412, 0.527)
Presenting symptomatically with stage D	0.85	0.82 (0.611, 0.989)
FOBT sensitivity to polyps	0.05 (0.00, 0.10)[a]	0.078 (0.0615, 0.1270)
FOBT sensitivity to CRC	0.4058 (0.30, 0.50)[a]	0.354 (0.3278, 0.3923)
FOBT specificity	98.5 (0.97, 0.99)[a]	0.9945 (0.9940, 0.9952)

CRC, colorectal cancer; FOBt, fecal blood occult test.
[a]Estimated from literature and expert opinion rather than within calibration.

REFERENCES

1. Tappenden P, Eggington S, Nixon R, Karnon J, Sakai H, Chilcott J. Colorectal Cancer Screening Options Appraisal. Report to the English Bowel Cancer Screening Working Group. Sheffield, UK: University of Sheffield; 2004.

2. Karnon J, Goyder E, Tappenden P, et al. A review and critique of modelling in prioritising and designing screening programmes. Health Technol Assess. 2007;11(52):iii–xi, 1.

3. Kim JJ, Kuntz KM, Stout NK, et al. Multiparameter calibration of a natural history model of cervical cancer. Am J Epidemiol. 2007;166:137–50.

4. Blower SM, Koelle K, Kirschner DE, Mills J. Live attenuated HIV vaccines: predicting the tradeoff between efficacy and safety. Proc Natl Acad Sci U S A. 2001;98:3618–23.

5. Harper PR, Jones SK. Mathematical models for the early detection and treatment of colorectal cancer. Health Care Manag Sci. 2005;8:101–9.

6. The Cancer Incidence and Surveillance Network (CISNET) website. www.cisnet.cancer.gov/colorectal/profiles.html

7. Loeve F, Boer R, van Oortmarssen GJ, van Ballegooijen, Habbema JD. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res. 1999;32:13–33.

8. Rutter M, Miglioretti DL, Savarino JE. Bayesian calibration of microsimulation models. J Am Stat Assoc. 2009;104:1338–50.

9. Higdon D, Kennedy M, Cavendish J, Cafeo J, Ryne RD. Combining field data and computer simulations for calibration and prediction. SIAM J Sci Comput. 2004;26:448–66.

10. Goldstein M, Rougier J. Bayes linear calibrated prediction for complex systems. J Am Stat Assoc. 2006;101(475):1132–43.

11. Spiegelhalter DJ. Incorporating Bayesian ideas into healthcare evaluation. Stat Sci. 2004;19:156–74.

12. National Institute for Health and Clinical Excellence (NICE). Guide to the Methods of Technology Appraisal. 2008. Available from: http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf

13. Cooper NJ, Sutton AJ, Abrams KR, Wailoo A, Turner DA, Nicholson KG. Effectiveness of neuraminidase inhibitors in treatment and prevention of influenza A and B: systematic review and meta-analyses of randomised controlled trials. Br Med J. 2003;326:1235.

14. Welton NJ, Ades AE. Estimation of Markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. Med Decis Making. 2005;25:633–45.

15. Tappenden P, Chilcott J, Eggington S, Patnick J, Sakai H, Karnon J. Option appraisal of population-based colorectal cancer screening programmes in England. Gut. 2007;56:677–84.

16. NHS Bowel Cancer Screening Programme. 2009. Available from: www.cancerscreening.nhs.uk/bowel/

17. Sharp L, Céilleachair A, Comber H, et al. Options for a Population-Based Colorectal Cancer Screening Programme in Ireland: A Health Technology Assessment. Cork, Ireland: Health Information & Quality Authority; 2009.

18. Nickerson C. Data from First Round of Bowel Cancer Screening in England. Sheffield, UK: NHS Cancer Screening Programmes; 2009.

19. Atkin WS, Saunders BP. Surveillance guidelines after removal of colorectal adenomatous polyps. Gut. 2002;51 Suppl 5:V6-V9.

20. Cotton S, Sharp L, Little J. The adenoma-carcinoma sequence and prospects for the prevention of colorectal neoplasia. Crit Rev Oncog. 1996;7:293–342.

21. Roberts GO, Gelman A, Gilks WR. Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann Appl Probab. 1997;7:110–20.

22. Plummer M, Best N, Cowles K, Vines K. CODA: Output analysis and diagnostics for MCMC. R package version 0.13-4. R News. 2006;6:7–11.

23. Incidence Data and Relative Survival Rates Calculated from the National Cancer Data Repository. Leeds, UK: Northern & Yorkshire Cancer Registry & Information Service and the National Cancer Intelligence Network; 2009.

24. United Kingdom Association of Cancer Registries (UKACR). Cancer Registry Population Denominators. Newport, South Wales: Office for National Statistics; 2009.

25. Munkholm P. Review article: the incidence and prevalence of colorectal cancer in inflammatory bowel disease. Aliment Pharmacol Ther. 2003;18(suppl 2):1–5.

26. Pendergrass CJ, Edelstein DL, Hylind LM, et al. Occurrence of colorectal adenomas in younger adults: an epidemiologic necropsy study. Clin Gastroenterol Hepatol. 2008;6:1011–5.

27. Eide TJ, Stalsberg H. Polyps of the large intestine in northern Norway. Cancer. 1978;42:2839–48.

28. Rickert RR, Auerbach O, Garfinkel L, Hammond EC, Frasca JM. Adenomatous lesions of the large bowel: an autopsy survey. Cancer. 1979;43:1847–57.

29. Williams AR, Balasooriya BA, Day DW. Polyps and cancer of the large bowel: a necropsy study in Liverpool. Gut. 1982;23:835–42.

30. Blatt LJ. Polyps of the colon and rectum: incidence and distribution. Dis Colon Rectum. 1961;4:277–82.

31. Vatn MH, Stalsberg H. The prevalence of polyps of the large intestine in Oslo: an autopsy study. Cancer. 1982;49:819–25.

32. Arminski TC, McLean DW. Incidence and distribution of adenomatous polyps of the colon and rectum based on 1,000 autopsy examinations. Dis Colon Rectum. 1964;7:249–61.

33. Office of National Statistics. Interim Life Tables 2005-07. 2009. Available from: www.statistics.gov.uk/StatBase/Product.asp?vlnk=14459

34. Burch JA, Soares-Weiser K, St John DJ, et al. Diagnostic accuracy of faecal occult blood tests used in screening for colorectal cancer: a systematic review. J Med Screen. 2007;14:132–7.

35. Weller D, Coleman D, Robertson R, et al. The UK colorectal cancer screening pilot: results of the second round of screening in England. Br J Cancer. 2007;97:1601–5.

36. Tappenden P, Eggington S, Nixon R, Chilcott J, Sakai H, Karnon J. Colorectal cancer screening options appraisal. Report to the English Bowel Cancer Screening Working Group. 2004. Available from: http://www.cancerscreening.nhs.uk/bowel/scharr.pdf

37. Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. Am Stat. 1995;49:327–35.

38. Cowles MK, Carlin BP. Markov chain Monte Carlo convergence diagnostics: a comparative review. J Am Stat Assoc. 1996;91:883–904.

39. Brooks SP, Roberts GO. Convergence assessment techniques for Markov chain Monte Carlo. Stat Comput. 1998;8:319–35.