Received: 28 January 2021 | Revised: 28 May 2021 | Accepted: 12 July 2021

DOI: 10.1002/hec.4408

RESEARCH ARTICLE

Health Economics WILEY

Flexible Bayesian longitudinal models for cost-effectiveness analyses with informative missing data

Alexina J. Mason[1] | **Manuel Gomes**[2] | **James Carpenter**[3,4] | **Richard Grieve**[1]

[1]Department of Health Services Research and Policy, LSHTM, University of London, London, UK

[2]Department of Applied Health Research, University College London, London, UK

[3]Department of Medical Statistics, LSHTM, University of London, UK

[4]MRC Clinical Trials Unit at UCL, London, UK

Correspondence
Alexina J. Mason, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK.
Email: alexina.mason@lshtm.ac.uk

Funding information
National Institute for Health Research, Grant/Award Number: SRF-2013-06-016; Medical Research Council, Grant/Award Numbers: MC_UU_ 12023/21, MC_ UU_ 12023/29

Abstract

Cost-effectiveness analyses (CEA) are recommended to include sensitivity analyses which make a range of contextually plausible assumptions about missing data. However, with longitudinal data on, for example, patients' health-related quality of life (HRQoL), the missingness patterns can be complicated because data are often missing both at specific timepoints (interim missingness) and following loss to follow-up. Methods to handle these complex missing data patterns have not been developed for CEA, and must recognize that data may be missing not at random, while accommodating both the correlation between costs and health outcomes and the non-normal distribution of these endpoints. We develop flexible Bayesian longitudinal models that allow the impact of interim missingness and loss to follow-up to be disentangled. This modeling framework enables studies to undertake sensitivity analyses according to various contextually plausible missing data mechanisms, jointly model costs and outcomes using appropriate distributions, and recognize the correlation among these endpoints over time. We exemplify these models in the REFLUX study in which 52% of participants had HRQoL data missing for at least one timepoint over the 5-year follow-up period. We provide guidance for sensitivity analyses and accompanying code to help future studies handle these complex forms of missing data.

KEYWORDS
Bayesian analysis, cost-effectiveness analysis, missing not at random, selection model, sensitivity analysis

1 | INTRODUCTION

International methods guidance for cost-effectiveness analyses (CEA) requires evidence about treatment effectiveness from well-designed randomized controlled trials (RCTs) (Sanders et al., 2016). Many CEA use RCT evidence about patients' health-related quality of life (HRQoL), measured at regular timepoints during the trial follow-up period. A common problem is that some of these longitudinal data are missing, as patients are lost to follow-up, or fail to complete the requisite questionnaires at each timepoint (Gabrio et al., 2017; Leurent et al., 2018a). Methods guidance requires the

study to consider the uncertainty pertaining to the missing data mechanism (Faria et al., 2014; Leurent et al., 2018b). CEA often assume that the missingness only depends on the observed data, in that the data are "missing at random" (MAR) (Gabrio et al., 2017; Leurent et al., 2018a). However, in many settings, missing data may depend on outcomes that are unobserved, for example, the patients' health status, and it is more reasonable to assume the data are "missing not at random" (MNAR) (Leurent et al., 2020; Mason et al., 2018). The "true" underlying missing data mechanism cannot be verified from the data at hand, and hence CEA are recommended to report sensitivity analyses according to alternative assumptions about missing data (Faria et al., 2014; Leurent et al., 2018b, 2020; Mason et al., 2018).

CEA with longitudinal data tend to have complex patterns of missing data which take several forms (Faria et al., 2014; Gabrio et al., 2020). For example, participants may be lost to follow-up, so that no further outcome data are available for that individual, or they may remain within the study, but fail to provide complete data at particular timepoints within the follow-up period (interim missingness). The approach to the missing data should then recognize that the reasons for loss to follow-up versus interim missingness may be different. Previous approaches to handling MNAR data in CEA have focused on pattern-mixture models and not considered these different forms of missing data within the longitudinal setting (Faria et al., 2014; Leurent et al., 2018b, 2020; Mason et al., 2018). Pattern-mixture models formulate the MNAR problem in terms of different distributions between missing and observed data. However, as these studies recognized, pattern-mixture models are less attractive to handle MNAR in longitudinal studies. For example, such forms of pattern-mixture model require strong assumptions about the differences between the observed and missing data distributions (sensitivity parameters) for each timepoint and do not readily allow the analyst to make plausible assumptions about the different forms of missing data.

Selection models offer an appealing approach to formulating the requisite sensitivity analyses in studies faced with different forms of missing data across multiple timepoints, and they have been applied within simple settings in comparative effectiveness research (Daniels & Hogan, 2008; Mason et al., 2012; Molenberghs et al., 2015). However, CEA raises additional challenges for the application of selection models for handling the missing data. First, costs and health outcomes tend to be correlated and need to be modeled jointly (Grieve et al., 2010; Nixon & Thompson, 2005; O'Hagan & Stevens, 2001). Second, CEA endpoints tend to have non-normal distributions, which complicates the MNAR modeling. In particular, considerable attention in the health econometrics literature has been given to developing models that recognize that HRQoL are left-skewed with spikes at 1, but these models have not been extended to common settings with missing data (Basu & Manca, 2012; Gomes et al., 2019; Hernandez-Alava et al., 2012). Modeling CEA endpoints according to plausible distributional assumptions is important in selection approaches because these directly identify the distribution of the unobserved values conditional on the observed data. Hence, currently available methods do not address fundamental concerns that arise when using longitudinal data in CEA.

Bayesian methods can help CEA provide evidence for directly informing decision-making, while allowing for complexities, such as the correlations between costs and health outcomes, the longitudinal structure in the data, and the need to make appropriate distributional assumptions (Baio, 2013; Lambert et al., 2008; Nixon & Thompson, 2005). Bayesian frameworks for addressing missing data in CEA have been proposed, but do not address the common challenges that arise with longitudinal data. The aim of this paper is to develop a Bayesian approach to CEA with longitudinal data, which uses selection models to make different, plausible assumptions about the missing data mechanism. This approach to sensitivity analyses is flexible, in that it can recognize different reasons for the missing outcome data at each timepoint, specify appropriate distributional assumptions for the costs and outcomes, and acknowledge the correlation between the endpoints. We exemplify our approach by re-analyzing the REFLUX study, which has a substantial proportion of patients (over 50%) with HRQoL data missing for at least one timepoint over the 5-year follow-up (Grant et al., 2013).

The remainder of this paper is structured as follows. Next we give more details on REFLUX, our motivating example that illustrates the challenges faced by CEA of longitudinal studies with MNAR data, and then we describe the proposed selection model framework. We then describe the development of the models through the re-analysis of the REFLUX case study, and present the results. The discussion highlights the main strengths and limitations of the proposed selection model approach and discusses some avenues for further research. Software code (using R and JAGS) is provided as supplementary material.

2 | METHODS

2.1 | Motivating example: the REFLUX trial

This CEA used information from an RCT in which 357 patients with moderately severe gastroesophageal reflux disease recruited from 21 hospitals were randomly assigned to laparoscopic surgery (LS) ($n = 178$), or medical management (MM) ($n = 179$) (Grant et al., 2013). QALYs were calculated from responses to the EQ-5D-3L questionnaire administered at baseline, 3 months and annually for 5 years post-randomization. Annual costs were collected and combined with QALYs to report estimates of relative cost-effectiveness over 5 years. In the base case analyses presented in the primary study publications, missing data were addressed using multiple imputation assuming MAR (Grant et al., 2013). The authors also presented a complete case analysis (CCA), and an MNAR sensitivity analysis that assumed patients with missing data at a particular follow-up timepoint had relatively low HRQoL, or high costs. The base case analysis and CCA suggested that LS was cost-effective relative to MM. The findings of the MNAR sensitivity analyses were somewhat mixed; the results were insensitive to alternative assumptions about the missing costs, but appeared less robust to different assumptions about the missing HRQoL data. In particular, the authors state:

> the cost-effectiveness of surgery is highly sensitive if it is assumed that surgery-allocated patients with missing data experience lower HRQoL than patients with complete data (Grant et al., 2013, p. 75).

While the REFLUX study did, therefore, consider the implications of missing data for the study conclusions, more flexible analytical approaches are required to address several related challenges that commonly occur with longitudinal HRQoL data. First, across the 5-year follow-up period, there are different forms of missing HRQoL data: 19.3% of patients have interim missing data, 24.9% loss to follow-up, and 7.6% have interim missingness, and then loss to follow-up. Figure 1 provides a more detailed representation of the missing HRQoL data patterns, and compares study arms. Second, LS is a "one-off" intervention, whereas MM could be provided throughout the follow-up period, and so the reasons for missing HRQoL data are likely to differ by treatment strategy. Third, the alternative forms of missing data may arise for different reasons; patients may be more inclined to "drop-out" following a large change in health status, whereas "interim" missingness could be "uninformative", that is unrelated to health status, or to reflect a temporary change in health or circumstances. Fourth, neither HRQoL nor costs are normally distributed (Figure 2). Fifth, total QALYs and total costs are correlated (correlation coefficient of −0.42 for the group assigned to MM, −0.07 for LS). Sixth, the study faced the common challenge of crossover, in that 67 (37.6%) patients randomized to LS received MM, and 10 (5.6%) patients randomized to MM received LS.

Motivated by these common concerns for CEA that use longitudinal data, we now propose a flexible Bayesian approach to handling missing data within the longitudinal setting.

2.2 | Proposed approach

2.2.1 | Bayesian selection model overview

We build on previous Bayesian methods for CEA (Baio, 2013, 2014; Gabrio et al., 2019, 2020; Grieve et al., 2010; Nixon & Thompson, 2005; O'Hagan & Stevens, 2001; Thompson & Nixon, 2005). We propose a Bayesian longitudinal selection model, which contains sub-models to handle the important complexities raised by missing data within the longitudinal setting. This approach harnesses the computational power and flexibility of Markov Chain Monte Carlo methods to undertake analyses that make different assumptions in addressing these complexities, within a single modeling framework. This model estimates a substantive model for the CEA endpoints (analysis model) and a model for the missingness (missingness model). Figure 3 shows the links between the various sub-models.

For ease of implementation, the joint analysis model for the health outcome and costs is specified as a marginal model for the health outcome and a conditional model for the costs. The three sub-models shown with a solid outline are fundamental for the CEA, but not all those with a dashed outline are necessarily required. This model could be further extended to allow for MNAR covariate missingness by adding a covariate missingness model.

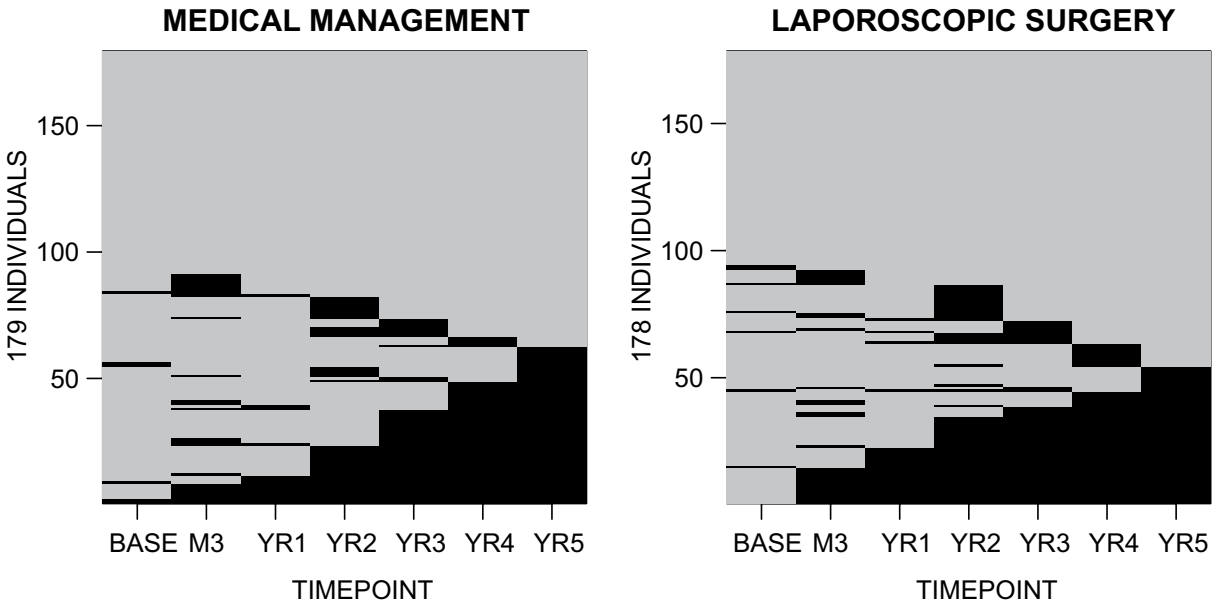

FIGURE 1 Pattern of missing health-related quality of life (HRQoL) by treatment arm. Black shading represents missing HRQoL for individuals (vertical axis) by timepoint (horizontal axis); gray shading represents observed HRQoL

2.2.2 | Strategy for developing component sub-models required by the REFLUX study

We now draw on the REFLUX case study to exemplify the steps required to build a Bayesian selection model with appropriate complexity for CEA based on data with longitudinal structure, and how to explore the impact of different model choices on the results. Figure 4 provides an overview.

2.3 | Step 1: select longitudinal analysis sub-model using complete cases

To further simplify the model building task, initially we work with each endpoint separately using available cases (172 for HRQoL and 191 for costs), and then assemble into a single analysis model. We consider three distributional assumptions for the HRQoL and cost endpoints: normal, gamma, and hurdle models (Grieve et al., 2010). Other options for HRQoL, such as a scaled beta or mixture models (Basu & Manca, 2012; Hernandez-Alava et al., 2012), could be included if exploratory data plots suggest that these are more appropriate. As LS is a "one-off" intervention whereas MM is an ongoing treatment strategy the trajectory of HRQoL over time may differ, and so each analysis model is parameterised separately according to treatment arm. However, to simplify notation, we suppress the treatment subscript "tr" in the model descriptions that follow.

The level of cross-overs rises substantially when individuals with partially observed data are also considered. Overall, 67 (37.6%) patients randomized to LS did not receive surgery, and 10 (5.6%) patients randomized to MM crossed over to receive surgery. To recognize that the distribution of costs and HRQoL reflected treatment received (cf. Figure 2d, which shows patients with total costs below £2000, who did not receive surgery), we analyze data from patients according to the treatment they actually received, but to address the decision problem of interest, we use the predictions from this analysis for patients as randomized, in line with an intention-to-treat (ITT) analysis (see Section 2.6).

2.3.1 | HRQoL analysis model

The REFLUX CEA required the study to report effect of randomized treatment on HRQoL each year, and then over the 5-year follow-up. We therefore chose to model HRQoL at each timepoint rather than aggregated. For simplicity, we initially ignored the multi-level structure in the data and fitted a model with each distributional assumption under consideration, incorporating the minimization covariates (age, BMI, and sex), baseline HRQoL, and time fixed effects.

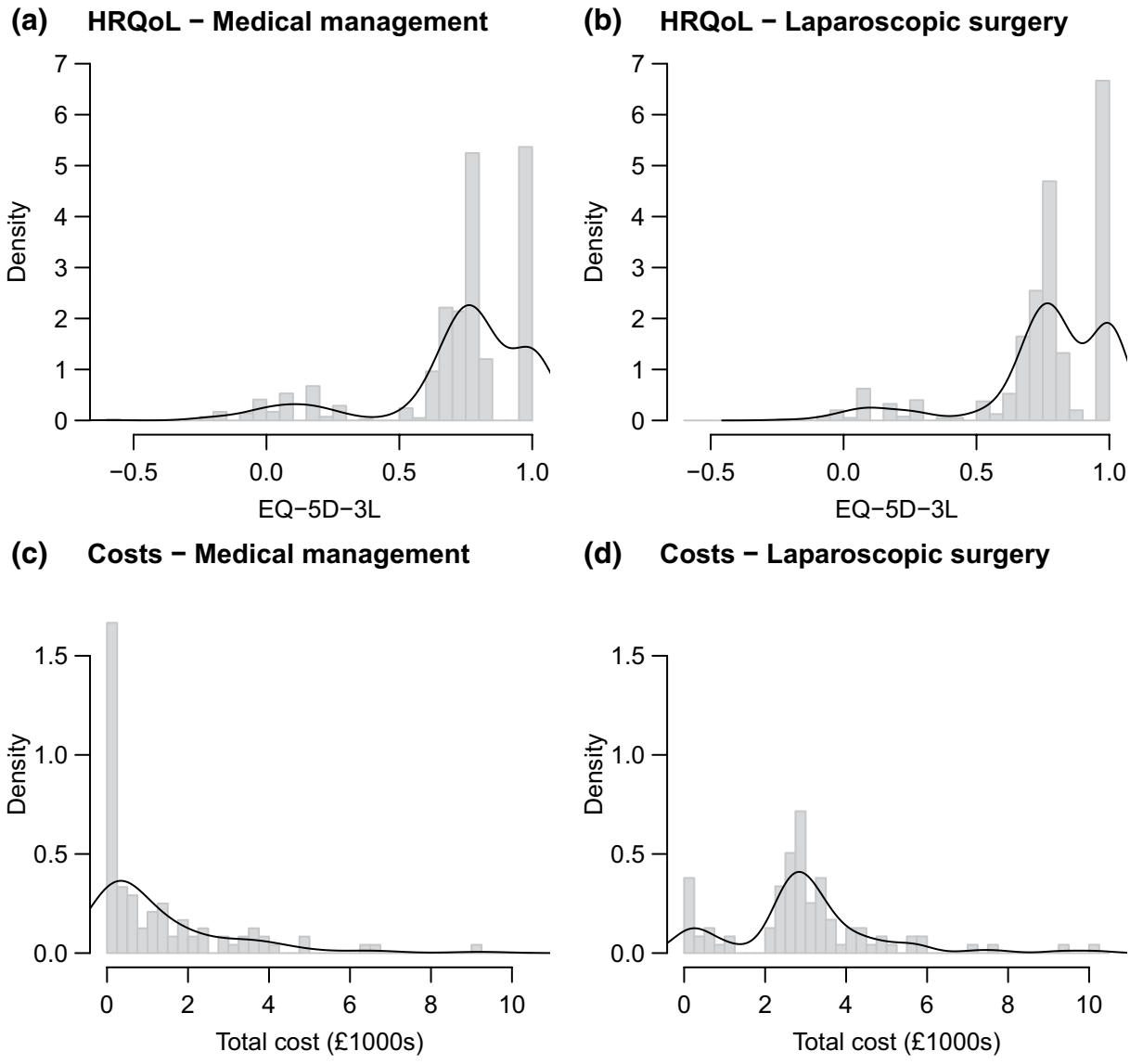

FIGURE 2 Distribution of health-related quality of life (HRQoL) (top) and costs (bottom) by treatment arm. An estimated kernal density has been superimposed on each histogram. HRQoL is across all timepoints (3 months and years 1–5)

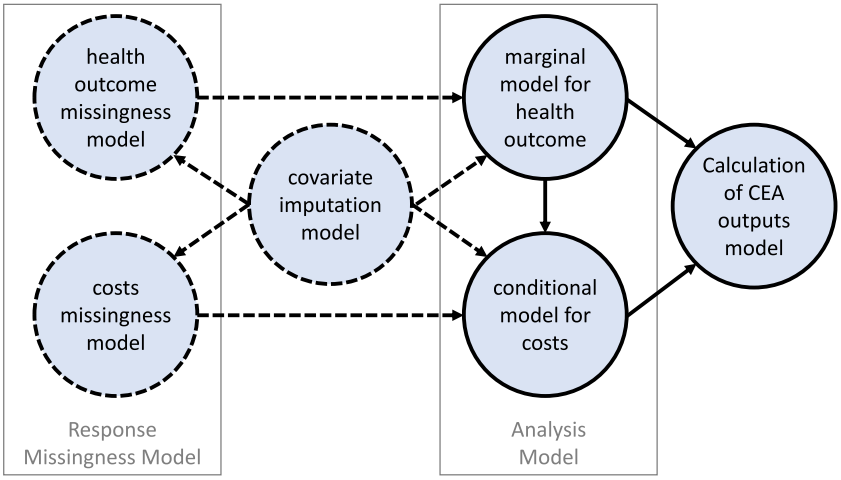

FIGURE 3 Schematic diagram of a typical Bayesian joint model for cost-effectiveness analysis. The sub-models shown with a solid outline will always be required, the requirement for those with a dashed outline depends on which variables have missing values and the assumptions about the missingness mechanism [Colour figure can be viewed at wileyonlinelibrary.com]

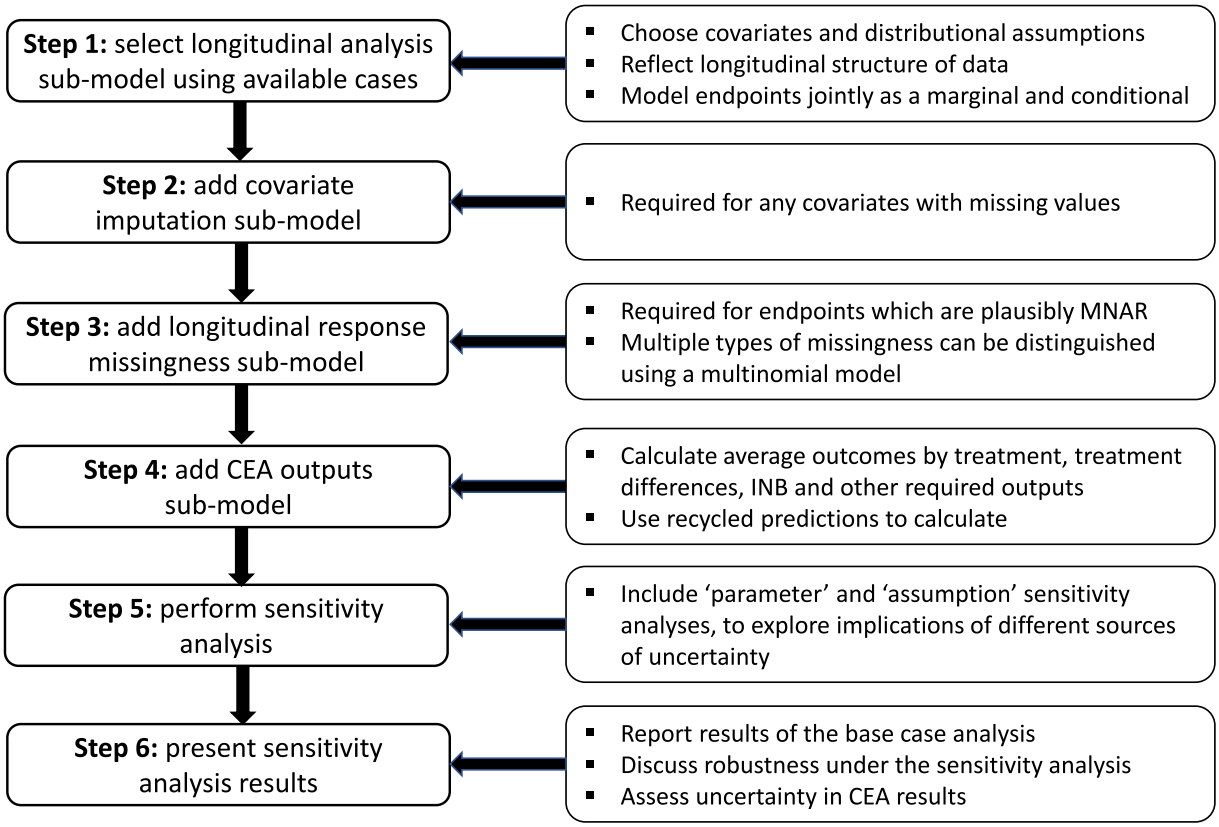

FIGURE 4 Modeling strategy for CEA based on longitudinal data. CEA, cost-effectiveness analysis
[Colour figure can be viewed at wileyonlinelibrary.com]

As the gamma distribution is restricted to positive values, we used HRQoL decrement ($1 - \text{HRQoL} + \epsilon$, where $\epsilon = 0.0001$ ensured positivity) as the health outcome for all models. Accordingly, for the hurdle option, we specified the hurdle at 0 and a gamma model for the non-zeros.

Next we added normally distributed patient random intercepts to each model to account for the multi-level structure. For the hurdle model, these were incorporated into the non-zeros part of the model and not the hurdle. The normal and hurdle models ran successfully, but convergence problems were encountered for the gamma model suggesting data and model incompatibility.

Overall fit can be compared between models by using the deviance information criteria (DIC) proposed by Spiegelhalter et al. (2002), with lower values suggesting a better fit. However, the DIC automatically produced by JAGS for the hurdle model is not directly comparable with the other models because the hurdle is also modeled. The DIC showed that the random effects improved the fit of both the normal and the hurdle models. We also examined specific aspects of model fit using residuals plots and posterior predictions (see Text S1 for examples).

Based on model fit, and as suggested by the exploratory data plots which show clear spikes at 1 in the HRQoL data (Figure 2), we chose the hurdle model with patient random intercepts for both treatments. Separate for each treatment arm, the full specification is as follows:

$$\begin{aligned} h_{it} &\sim \text{Bernoulli}(p_{it}) \\ \text{logit}(p_{it}) &= \gamma_t + \omega_0 bq_i + \omega_1 bmi_i + \omega_2 age_i + \omega_3 sex_i \\ q_{it} \mid (h_{it} = 1) &= 0 \\ q_{it} \mid (h_{it} = 0) &\sim \text{Gamma}(shape.q, shape.q/\mu.q_{it}) \\ \log(\mu.q_{it}) &= \alpha_t + \theta_i + \beta_0 bq_i + \beta_1 bmi_i + \beta_2 age_i + \beta_3 sex_i \\ \theta_i &\sim \text{Normal}(\theta.\mu, \theta.\sigma^2) \end{aligned} \tag{1}$$

where q_{it} is the HRQoL decrement for patient i at time t, bq_i denotes baseline HRQoL decrement for patient i and θ_i are patient random intercepts. Minimally informative priors are placed on the unknown parameters (see Appendix for details).

2.3.2 | Cost analysis model

For the REFLUX study, like many CEA, interest is more on the effect of treatment assignment on costs over the full-time horizon, rather than for each year. We therefore chose to model costs at the aggregate 5-year level, allowing us to demonstrate an alternative way of incorporating partially observed values. According to DIC, our chosen model is a gamma for both treatments, including the covariates BMI, age, and sex.

2.3.3 | Joint analysis model for HRQoL and total costs

We now combine the component sub-models into the proposed joint model. For each treatment arm, we specify the conditional cost model as follows:

$$\begin{aligned} c_i &\sim \text{Gamma}(shape.c, shape.c / \nu.c_i) \\ \log(\nu.c_i) &= \zeta_0 + \zeta_1 bmi_i + \zeta_2 age_i + \zeta_3 sex_i + \xi(Q_i - \mu.Q_i) \end{aligned} \tag{2}$$

where c_i is the aggregated 5-year costs in GBP for patient i. We have specified the second parameter of the gamma function (rate) in terms of the shape parameter and the conditional 5-year costs mean for individual i, $\nu.c_i$. Q_i and $\mu.Q_i$ are the estimated 5-year QALYs and estimated 5-year QALY mean, respectively, for individual i. See Appendix for prior and other implementation details. The HRQoLs at each timepoint are combined into 5-year QALYs by linear interpolation, according to the "area under the curve" method, as follows:

$$Q_i = 0.5(1-q_{i1}) + 0.875(1-q_{i2}) + \sum_{t=3}^{5}(1-q_{it}) + 0.5(1-q_{i6}) \tag{3}$$

Similarly, $\mu.Q_i$ can be estimated by combining predicted values of q_i at each timepoint ($pred.q_{it}$), where

$$\begin{aligned} pred.h_{it} &\sim \text{Bernoulli}(p_{it}) \\ pred.q_{it} &= (1 - pred.h_{it}) \times \mu.q_{it} \end{aligned} \tag{4}$$

The marginal costs can be recovered using

$$\mu.c_i = \exp(\zeta_0 + \zeta_1 bmi_i + \zeta_2 age_i + \zeta_3 sex_i) \tag{5}$$

So far, we have only fitted this analysis model to patients with available data. Without making any changes, we can also include data from patients with partially observed responses provided their covariates are fully observed. For patients with HRQoL at some but not all timepoints, the specification of the disaggregated HRQoL model directly incorporates the values from observed timepoints and imputes the values that are missing. To incorporate partially observed cost information, we place a lower limit (calculated as the sum of the observed costs) on the gamma distribution for the missing aggregate 5-year costs. Without the addition of a response missingness model, the missing values are drawn from the posterior distribution assuming MAR.

2.4 | Step 2: add covariate imputation sub-model

An analysis model will run with missing responses, but not with missing covariates. So, the next step is to specify a covariate imputation model to impute any missing covariates. For REFLUX, BMI, age, and sex are all fully observed, but baseline HRQoL has 13 (3.6%) missing values. Given the low level of missingness, we model baseline HRQoL decrement using a Gamma distribution, that is,

$$bq_i \sim \text{Gamma}(shape.bq, rate.bq)T(,1.2) \tag{6}$$

where $T()$ imposes an upper bound of 1.2 on the HRQoL decrements to restrict the imputations to viable values.

2.5 | Step 3: add longitudinal response missingness sub-model

For REFLUX, it seems likely that the probability of a patient providing their HRQoL at a particular timepoint is related to their health status at that time, so we add a heath outcome missingness model to explore different MNAR assumptions. By contrast, it is more reasonable to assume that costs are MAR as missingness is more likely to reflect administrative reasons (e.g., missing case notes) rather than a patient's unobserved health status, and so we do not specify a cost outcome missingness model. If investigators consider that costs may also be MNAR, the response missingness model can be extended to encompass both CEA endpoints (Figure 3).

When there is a single type of missingness, a response missingness model can be specified as a logistic model for a binary missing value indicator, m_i (0 = observed, 1 = missing) for individual i. However, as with REFLUX, typically both interim missingness and loss to follow-up occur in longitudinal studies. To distinguish between multiple types of missingness, this model can be extended by specifying a multinomial logistic model for a categorical missing value indicator.

For our illustration, we incorporate covariates (age, BMI, and sex), time fixed effects, the immediate previous HRQoL, *previous.q* (baseline HRQoL is used for the first timepoint), and change from previous HRQoL, *change.q*. It is the inclusion of the possibly unobserved *change.q* that changes the assumption about the missing HRQoL from MAR to MNAR, and provides the link with the analysis model. The extent to which the missingness mechanism is assumed to depart from MAR, is captured by the parameters of *change.q*, $\boldsymbol{\lambda}$. Daniels and Hogan (2008) define a *sensitivity parameter* to be a parameter that is completely non-identified by the data. $\boldsymbol{\lambda}$ are not sensitivity parameters in this strict sense, as their estimation draws on the parametric assumptions in the analysis model and response missingness model (Mason et al., 2012)—Daniels and Hogan (2008), Section 8.3.2, provides clear examples of how this works. Indeed, selection models cannot be factorized into identifiable and non-identifiable parts. However, estimation of $\boldsymbol{\lambda}$ type parameters can be difficult as it is reliant on limited information from assumptions about other parts of the model and informative priors are recommended (Mason et al., 2012). Therefore, we recommend giving $\boldsymbol{\lambda}$ point priors and explore the sensitivity of the CEA outputs to different values.

As in the analysis model, all the parameters are allowed to differ by treatment arm. Consequently, point priors are required for four λ parameters: MM arm interim missing, MM arm loss to follow-up, LS arm interim missing, and LS arm loss to follow-up, with the choice informed by substantive knowledge about each intervention.

Setting m_{it} to be a three-category missing value indicator for q_{it} for patient i at time t (1 = observed, 2 = interim missing, 3 = loss to follow-up), the full specification of this multinomial logistic sub-model is as follows (suppressing the treatment subscript, tr, to simplify notation):

$$\begin{aligned} \boldsymbol{count}_{it} &\sim \text{Multinomial}(\boldsymbol{s}_{it}, 1) \\ \phi_{it,1} &= 1 \\ \log(\phi_{it,r}) &= \kappa 0_{t,r} + \kappa_{r1} bmi_i + \boldsymbol{\kappa}_{r2} age_i + \kappa_{r3} sex_i + \kappa_{r4} previous.q_{it} + \lambda_r change.q_{it}; \quad \text{r} = 2,3 \\ s_{it,r} &= \frac{\phi_{it,r}}{\sum_{z=1}^{3} \phi_{it,z}} \end{aligned} \tag{7}$$

where r indicates the missingness category, and ***count*** is a vector with $count_{it,r}$ set to 1 if $m_{it,r} = r$ and 0 otherwise. See Appendix for prior specifications. Since this sub-model is dependent on partially observed covariates, it also links with the covariate imputation model.

2.6 | Step 4: add CEA outputs sub-model

We report incremental cost-effectiveness using the incremental net monetary benefit (INB). The Bayesian approach allows the INB to be calculated from the posterior distribution of the parameter estimates through specifying a set of equations, which we call the CEA outputs sub-model (see Appendix for details). The uncertainty from estimating each sub-model is, therefore, propagated through to the posterior distribution of the INB, and can be encapsulated in the credible intervals around the INB estimates and other metrics such as the cost-effectiveness acceptability curve.

We use the method of *recycled predictions* which can accommodate GLMs with non-linear link functions in predicting the incremental effects (Basu & Manca, 2012; Glick et al., 2007). This method uses the fitted model to predict incremental effects using only the baseline covariates of the patients randomised into the trial, and proceeds as follows:

1. Use patient-level baseline covariates to predict outcome for all patients, assuming they are randomized to a particular treatment arm, for example usual care.
2. Analogously, predict outcome for all patients, assuming they are randomized to new treatment.
3. Calculate difference between the outcomes predicted in points 1 and 2 for each individual.
4. Incremental effects = mean of differences calculated in point 3.

To perform an ITT analysis, we predict the outcomes for cross-over patients according to the treatment they received in the trial for both points 1 and 2. Accordingly, randomization does not change their predictions, but it does tell us how to average them to obtain the ITT CEA estimate. This approach relies on a model that accurately captures the key features of the data, raising the importance of model choice.

2.7 | Step 5: perform sensitivity analysis

To illustrate, we investigate the eight sensitivity scenarios shown in Table 1. Here, point priors for the selection parameters λ_r in Equation (7) are: (i) "positive MNAR selection," a value of $0.69 = \log(2)$, which encodes an assumption of a twofold increase in the probability of being missing for a change of 1 unit on the HRQoL scale (conditional on other variables in the selection model); (ii) MAR, corresponding to a value of zero, and (iii) "negative MNAR selection," a value of $-0.69 = \log(1/2)$. Relative to MAR, "positive MNAR" selection leads to higher imputations for the missing components of HRQoL, and "negative MNAR" leads to lower imputations. We have chosen a twofold increase because, based on our experience, this is at the limits of plausibility. For a "live" trial, we recommend consulting experts familiar with the disease, patient population and treatments.

In Scenarios 1 and 2, for each treatment the two types of missingness are assumed to be caused by similar mechanisms, but the causes of the missingness are assumed to be different for the two treatments and have opposite effects. This type of situation has the greatest potential to lead to conclusion changing differences in the treatment effect compared with assuming MAR throughout. Scenarios 3–6 assume one type of missingness is MAR, but the other type is MNAR with the opposite effect on the two treatments. These situations will likely lead to smaller differences compared with an all MAR scenario. For Scenario 7, missingness is assumed to be associated with lower HRQoL for both missingness types and treatments, while Scenario 8 is the higher HRQoL equivalent. For Scenarios 7 and 8, any change in treatment differences will be due solely to differences in the rates of missingness between the treatments.

Figure 5 shows a posterior density strip (Jackson, 2008) of imputed HRQoL data for each of three scenarios for three patients, to demonstrate the considerable uncertainty in the imputations, and to provide insight into how the posterior distributions of these imputations shift according to the missing data assumptions. The patient in the left most panel has interim missing data at 2 and 4 years; the patient in the center panel drops out from 3 years onwards and the patient in the right panel has an interim missing value at 2 years before dropping out at 4 years. The hurdle model is clearly seen in the low posterior density (light color density) just below 1. As we might expect, for the patient receiving LS, the posterior distribution from imputation under Scenario 1 (red) gives greater probability to higher values than the posterior distribution from imputation under Scenario 2 (blue). The opposite is true for the two patients receiving MM. While we might expect the posterior mean of the distribution under MAR (black) to fall between the posterior means from the two MNAR scenarios, this is not always true because these models do not contain only the sensitivity parameters; the estimated values of other model parameters will change as they are estimated to obtain best fit.

To demonstrate Step 5 of our modeling strategy, we have focused on exploring sensitivity to the choice of the selection parameters. If relevant external information is available from historical sources or experts, then further "parameter sensitivity" analysis could incorporate informative priors on other model parameters. Additionally, in studies where the choice of distribution is less obvious, we also recommend carrying out an "assumption sensitivity" to explore alternative distribution assumptions.

2.8 | Step 6: present sensitivity analysis results

To demonstrate how our approach allows the comparison of scenarios at a disaggregate level, for analysis under complete cases only, the MAR assumption, and MNAR Scenarios 1 and 2, respectively, Table 2 shows details of mean posterior QALYs for each treatment, their difference between treatments and the probability that the difference favors LS for each

TABLE 1 Eight MNAR sensitivity scenarios

	Medical management		Laparoscopic surgery	
	Interim missing	**Loss to follow-up**	**Interim missing**	**Loss to follow-up**
Scenario 1:	↓	↓	↑	↑
Scenario 2:	↑	↑	↓	↓
Scenario 3:	—	↓	—	↑
Scenario 4:	—	↑	—	↓
Scenario 5:	↓	—	↑	—
Scenario 6:	↑	—	↓	—
Scenario 7:	↓	↓	↓	↓
Scenario 8:	↑	↑	↑	↑

Note: "↑" positive MNAR selection—imputed values higher than MAR; "—" MAR imputation; "↓" negative MNAR selection—imputed values lower than MAR.

Abbreviations: MAR, missing at random; MNAR, missing not at random.

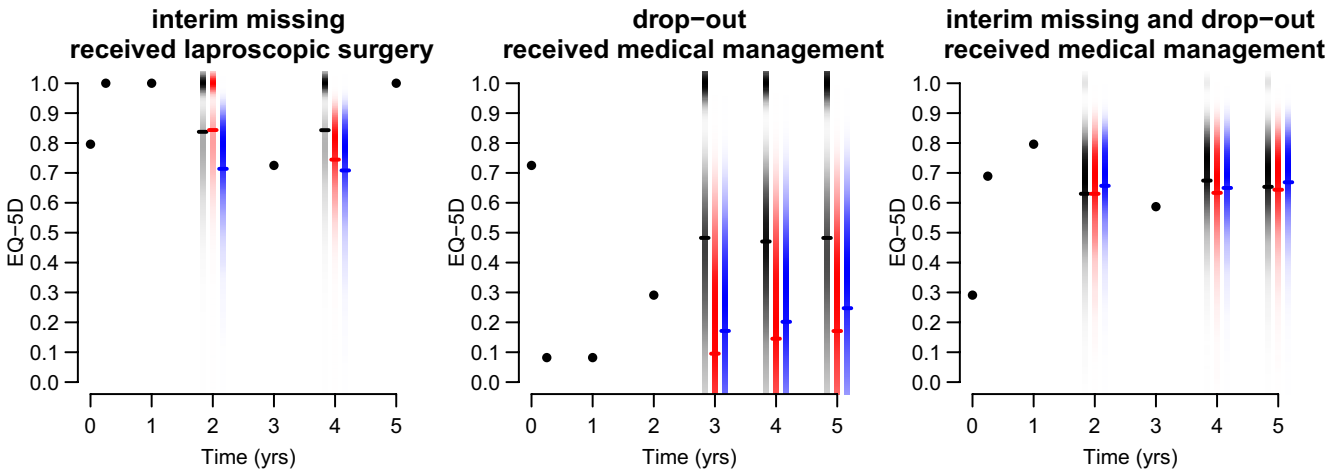

FIGURE 5 Observed and imputed data for three patients. In each panel, the closed black circles indicate observed data, and the colored strips indicate the posterior distribution of imputed data under: black: MAR; red: Table 1 Scenario 1, and blue: Table 1 Scenario 2. The probability density is represented by the color density (note the bimodal distributions for some scenarios), and the posterior mean is marked "–" [Colour figure can be viewed at wileyonlinelibrary.com]

year, and the sum of the QALY over the 5 years of follow-up ("5 year Total"). Comparison with the CCA analysis indicates where incorporating information from partially observed patients is particularly important. For all scenarios, there is a high probability that patients benefit from LS compared to MM throughout the 5 years, but these benefits reduce over time. The magnitude of the benefit varies between scenarios, with the benefits approximately doubling for MNAR Scenario 1 compared to MAR.

Table 3 summarizes costs and QALYs over the 5-year period for the eight MNAR scenarios, compared with MAR and CCA. Among the MNAR scenarios, MNAR1 and MNAR2 produce the highest and lowest QALY differences, respectively. The results for Scenarios MNAR3–MNAR6 reveal that in these data, the differences are driven by the assumptions about the loss to follow-up (MNAR3 is very close to MNAR1, and likewise MNAR4 to MNAR2), rather than the interim missingness (little difference between MNAR5 and MNAR6). The differences in the costs across the MNAR scenarios are small, since missing costs are always assumed to be MAR. The differences between the MAR and MNAR scenarios show the implications of the joint estimation of the analysis model and response missingness model.

Figure 6 shows INB, valuing quality-adjusted life year gains at 20,000 GBP per quality-adjusted life year, for the complete cases, analysis under MAR, and each of the eight MNAR scenarios in Table 1. As expected, comparing the MAR posterior distribution (shown as a density strip) and the 95% credible interval with those for CCA, reveals that incorporating extra information from the partially observed patients reduces uncertainty. However, allowing observed and

TABLE 2 QALY summary by year: CCA, MAR, and two MNAR scenarios

	Mean LS[a]	Mean MM[a]	Difference[a]	Prob[b]
CCA				
Year 1	0.80 (0.77,0.83)	0.72 (0.69,0.75)	0.08 (0.04,0.12)	1.000
Year 2	0.77 (0.73,0.80)	0.71 (0.67,0.74)	0.06 (0.01,0.10)	0.996
Year 3	0.75 (0.72,0.78)	0.71 (0.68,0.74)	0.04 (0.00,0.08)	0.976
Year 4	0.74 (0.70,0.76)	0.69 (0.67,0.72)	0.04 (0.00,0.08)	0.984
Year 5	0.70 (0.67,0.73)	0.67 (0.64,0.69)	0.04 (0.00,0.07)	0.962
5 year total	3.76 (3.64,3.87)	3.50 (3.38,3.61)	0.26 (0.10,0.41)	0.999
MAR				
Year 1	0.78 (0.75,0.81)	0.70 (0.67,0.72)	0.06 (0.03,0.09)	1.000
Year 2	0.75 (0.72,0.78)	0.68 (0.65,0.71)	0.05 (0.02,0.09)	0.999
Year 3	0.73 (0.70,0.76)	0.68 (0.65,0.70)	0.04 (0.01,0.08)	0.993
Year 4	0.71 (0.68,0.74)	0.67 (0.64,0.69)	0.03 (0.00,0.06)	0.985
Year 5	0.68 (0.65,0.71)	0.64 (0.62,0.67)	0.03 (0.00,0.06)	0.968
5 year total	3.66 (3.54,3.77)	3.36 (3.25,3.47)	0.22 (0.10,0.35)	1.000
MNAR scenario 1[c]				
Year 1	0.73 (0.70,0.76)	0.62 (0.57,0.66)	0.11 (0.07,0.16)	1.000
Year 2	0.69 (0.65,0.72)	0.59 (0.53,0.63)	0.10 (0.05,0.16)	1.000
Year 3	0.66 (0.62,0.69)	0.58 (0.52,0.62)	0.08 (0.03,0.14)	1.000
Year 4	0.64 (0.60,0.67)	0.57 (0.52,0.61)	0.07 (0.03,0.12)	0.999
Year 5	0.62 (0.58,0.65)	0.55 (0.50,0.59)	0.07 (0.02,0.12)	0.998
5 year total	3.34 (3.18,3.47)	2.90 (2.65,3.10)	0.44 (0.23,0.68)	1.000
MNAR scenario 2[c]				
Year 1	0.73 (0.70,0.76)	0.64 (0.60,0.68)	0.08 (0.04,0.13)	1.000
Year 2	0.68 (0.63,0.71)	0.61 (0.56,0.65)	0.06 (0.01,0.12)	0.992
Year 3	0.65 (0.61,0.69)	0.60 (0.56,0.64)	0.04 (−0.01,0.09)	0.956
Year 4	0.63 (0.59,0.67)	0.60 (0.55,0.63)	0.03 (−0.01,0.08)	0.917
Year 5	0.61 (0.57,0.64)	0.58 (0.54,0.61)	0.03 (−0.02,0.08)	0.901
5 year total	3.29 (3.11,3.44)	3.03 (2.82,3.21)	0.26 (0.05,0.47)	0.991

Abbreviations: CCA, complete case analysis; LS, laparoscopic surgery; MAR, missing at random; MNAR, missing not at random; MM, medical management.

[a]Posterior mean (95% credible interval).

[b]Probability favors LS.

[c]MNAR scenarios as defined in Table 1.

unobserved HRQoL to be systematically different increases uncertainty, as shown by the increased interval widths in the MNAR scenarios, compared with both MAR and CCA.

Intuitively, we expect the estimated INB to be higher compared to the MAR analysis in MNAR scenarios with positive MNAR selection for LS and negative MNAR selection for MM (MNAR1—both types of missingness; MNAR3—loss to follow-up only; MNAR5—interim missing only), as this will increase HRQoL differences between the two treatments. Figure 6 is consistent with this expectation, and also shows that it is the loss to follow-up rather than the interim missingness which predominantly drives the increased difference. For MNAR scenarios with negative MNAR selection for LS and positive MNAR selection for MM (MNAR2—both types of missingness; MNAR4—loss to follow-up only; MNAR6—interim missing only) we expect the reverse effect, but although each of these three scenarios shifts the INB posterior density to the left of its MNAR counterpart, their posterior means are higher than for MAR. This is because, as discussed in Section 2.7, these models do not contain pure sensitivity parameters, and there will be some balancing out as the estimated values of other model parameters are adjusted to obtain best fit. For Scenarios MNAR7 and MNAR8, the MNAR selection is in the same direction for both LS and MM, so given that the proportion of missing HRQoL values is reasonably balanced across treatment arm, we expect the INB posterior means to be similar to MAR. However, consistent with the results from other MNAR scenarios, these are higher due to model fitting adjustments in other parameters.

TABLE 3 Five-year summary: CCA, MAR, and MNAR sensitivity analyses

	QALYs			Costs (£1000s)		
	Mean LS[a]	Mean MM[a]	Difference[a]	Mean LS[a]	Mean MM[a]	Difference[a]
CCA	3.76 (3.64,3.87)	3.50 (3.38,3.61)	0.26 (0.10,0.41)	3.33 (3.06,3.63)	1.41 (1.03,2.05)	1.92 (1.24,2.41)
MAR	3.66 (3.54,3.77)	3.36 (3.25,3.47)	0.22 (0.10,0.35)	3.53 (3.27,3.82)	1.21 (0.93,1.60)	1.83 (1.46,2.16)
MNAR1	3.34 (3.18,3.47)	2.90 (2.65,3.10)	0.44 (0.23,0.68)	3.21 (2.97,3.47)	1.34 (1.05,1.74)	1.87 (1.49,2.21)
MNAR2	3.29 (3.11,3.44)	3.03 (2.82,3.21)	0.26 (0.05,0.47)	3.22 (2.98,3.49)	1.34 (1.05,1.74)	1.88 (1.50,2.23)
MNAR3	3.34 (3.18,3.47)	2.90 (2.65,3.10)	0.44 (0.23,0.67)	3.21 (2.98,3.47)	1.34 (1.05,1.75)	1.87 (1.48,2.21)
MNAR4	3.29 (3.12,3.44)	3.03 (2.83,3.21)	0.26 (0.05,0.47)	3.22 (2.98,3.49)	1.34 (1.05,1.74)	1.88 (1.50,2.23)
MNAR5	3.32 (3.15,3.46)	2.97 (2.74,3.15)	0.35 (0.15,0.58)	3.21 (2.98,3.47)	1.34 (1.05,1.73)	1.87 (1.49,2.21)
MNAR6	3.32 (3.15,3.46)	2.96 (2.73,3.15)	0.36 (0.15,0.58)	3.21 (2.97,3.48)	1.34 (1.05,1.73)	1.87 (1.49,2.22)
MNAR7	3.26 (3.08,3.41)	2.89 (2.65,3.10)	0.37 (0.14,0.61)	3.22 (2.98,3.48)	1.34 (1.05,1.75)	1.87 (1.48,2.21)
MNAR8	3.37 (3.22,3.50)	3.04 (2.82,3.22)	0.33 (0.13,0.53)	3.21 (2.97,3.46)	1.34 (1.04,1.72)	1.87 (1.50,2.21)

Abbreviations: CCA, complete case analysis; LS, laparoscopic surgery; MAR, missing at random; MM, medical management; MNAR, missing not at random.
[a]Posterior mean (95% credible interval).

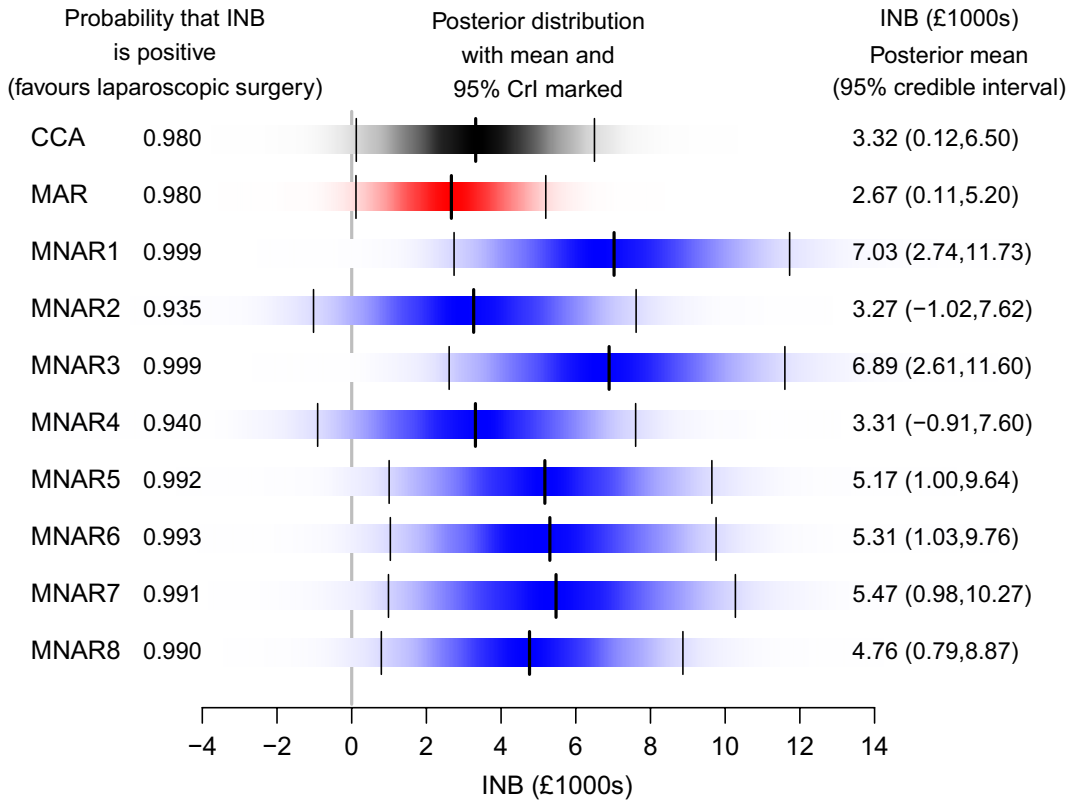

FIGURE 6 Comparison of incremental net benefit (INB). Each shaded rectangular strip shows the full posterior distribution of the incremental net benefits, valuing quality-adjusted life year gains at 20,000 GBP per quality-adjusted life year. The color density is proportional to the probability density, such that the strip is darkest at the maximum density and fades into the background at the minimum density. The posterior mean and 95% credible interval are marked. CCA, complete case analysis; CrI, credible interval; MAR, missing at random; MNAR, missing not at random [Colour figure can be viewed at wileyonlinelibrary.com]

In six of the MNAR scenarios, the probability that INB is positive is at least 99%, strongly suggesting that surgery is cost-effective. However, there is some sensitivity when missing values for patients receiving surgery are assumed lower than for other patients (MNAR2 and MNAR4), but the probabilities are still over 90% providing confidence that the findings in the primary analysis are relatively robust. Interest would likely focus on the plausibility of these scenarios in a policy decision discussion, and in a "live" CEA, further sensitivity analysis would probe these scenarios.

3 | DISCUSSION

This paper has developed and illustrated Bayesian selection models for CEA with informative missing data within longitudinal studies. This approach can be applied to undertake sensitivity analyses that make clearly defined, transparent assumptions. These flexible models allow the assumed missing data mechanism to differ by treatment strategy, but also according to whether the missing data reflects loss to follow-up or interim missing values. The approach also addresses typical challenges arising in CEA, such as the need to jointly model costs and outcomes, to handle non-normal distributions, and to accommodate non-compliance with the treatment assigned.

We illustrate how this approach can improve the interpretation of a study's results, by revisiting and extending the previous analyses of the REFLUX trial (Grant et al., 2013). In the original analysis, the authors did consider that data may be MNAR, specifically assuming that HRQoL may be lower in (a) all patients with missing data and (b) those randomized to surgery who had missing data. The authors concluded that the overall conclusion, that surgery was cost-effective, was somewhat sensitive to assuming lower HRQoL in the surgery arm alone. Our methodology permits a re-analysis with a more extensive range of MNAR scenarios (8), and finds that while the results are slightly more sensitive to assumptions about "loss to follow-up" versus "interim missingness", the conclusion that surgery is more cost-effective is robust to a wide range of alternative assumptions about the missing data mechanism. This approach can be applied directly in future studies to consider the impact of alternative missing data mechanisms on their conclusion, and in other settings, it may also be useful to distinguish the impact of "loss to follow-up" from "interim missingness". For example, the framework can be particularly useful in studies with differential loss to follow-up according to the comparison group, as might occur in evaluating a new versus old treatment for metastatic cancer. Alternatively, the approach may accommodate greater levels of interim missingness if the new treatment has a higher incidence of side effects, versus the standard of care. In these settings, this approach provides a framework for assessing the importance of alternative, realistic assumptions about the level of HRQoL for those patients with missing data, and the potential impact on the study's conclusions.

The proposed Bayesian selection models have important advantages compared to sensitivity analysis strategies in CEA that use pattern mixture models (Faria et al., 2014; Gabrio et al., 2020; Leurent et al., 2020; Mason et al., 2018), and build on Bayesian approaches to CEA that use RCT data (Baio, 2013; Gabrio et al., 2020; Lambert et al., 2008) and related research using Bayesian models to estimate HRQoL (Kharroubi et al., 2005, 2015, 2018). Practical advantages over pattern mixture models include (i) adopting the selection model approach enables simple conditional models to be added at each step, with some sub-models "discretionary" according to the setting; (ii) the approach can distinguish between "loss to follow-up" and "interim missingness" patterns by specifying a multinomial missing data model; (iii) the missing data and endpoint models can easily accommodate the longitudinal structure of the data, and (iv) the selection model approach requires relatively few sensitivity parameters, whereas the pattern mixture approach would require a distinct subset of models for each missing data pattern, implying many models within typical longitudinal settings. The fully Bayesian approach to selection modeling allows the uncertainty associated with the missing data to be fully propagated through the whole model, and is reflected in the final estimates of incremental cost-effectiveness.

Our example, the REFLUX study, is typical of many studies both in terms of the distribution of the costs and effects, and the longitudinal follow-up with interim missing data and loss to follow-up. Therefore, our approach has wide applicability. The breadth of applicability can be further expanded by noting that the components can be modified for CEA with different features. For example, the analyst can easily change the specification (e.g., distributions) of the models for analyzing HRQoL, for example, to incorporate beta-type models (Basu & Manca, 2012), or mixture models (Hernandez-Alava et al., 2012), or more flexible approaches to model cost data (Mihaylova et al., 2011). Also, often studies will have clinical outcomes that are correlated with the QALYs, costs, or missingness. These outcomes can be incorporated into the relevant sub-model, in the same way that we have incorporated baseline covariates, to improve the robustness of the missingness adjustments. While we focus on addressing MNAR in the health outcome, the missingness model could be developed for costs (or both endpoints). To encourage the uptake of the proposed methods, and help future studies tailor them to their needs, we provide accompanying software code to implement these models in R and JAGS (see Text S1).

Nevertheless, there are some limitations to the proposed implementation of the approach, which motivate areas for further research. First, we obtain the metrics of interest with the method of recycled predictions, in which the model predicts each endpoint for all patients for each treatment alternative. The potential drawback of this approach is in assuming the endpoint model is correctly specified. Here, the gamma-hurdle model fitted the observed data relatively well, but in other settings, it may be necessary to consider whether the results are robust to a wider set of model choices, or use observed outcome data (QALY, cost) whenever this is available. Second, the selection models were developed following publication of the primary results. Typically, we wish to pre-specify the model as part of the health economics analysis

plan (Thorn et al., 2020). For selection models this requires that the analyst specifies plausible values for the sensitivity parameters a priori. A natural approach would be to elicit plausible values from experts, building on the elicitation approaches developed in Mason et al. (2018). We believe this is more sensible than the "tipping point" approach, where parameters are typically moved from the base case (typically MAR) scenario until "conclusions change" and then a posterior judgment is made as to the plausibility of these relatively extreme scenarios. Thirdly, while the REFLUX study exemplified many concerns typical in CEA that use longitudinal data, it focuses on continuous endpoints, and did not consider time to event measures such as survival time.

The framework proposed can accommodate survival outcomes but would require substantial changes in the ways the models are parameterized. Here, the response missingness model would be replaced by a model for the time to drop out, allowing informative censoring. These models can capture key features of the observed data, but would need to make plausible predictions for the period beyond the observed data (Baio, 2020; Guyot et al., 2017; Rutherford et al., 2020). More generally, our proposed modeling strategy can be used for other non-continuous types of endpoints, but the model specification would require some adaptation.

In conclusion, this Bayesian selection modeling approach, has both the flexibility and robustness to be pre-specified for the majority of CEA analyses that use RCTs with longitudinal data. We provide annotated code to support its application in future studies.

ACKNOWLEDGMENTS

We thank Craig Ramsay, David Epstein, Rita Faria, and Mark Sculpher for access to the REFLUX data. This report is independent research partly supported by the National Institute for Health Research (Senior Research Fellowship, RG, SRF-2013-06-016). JC is supported by the Medical Research Council, grant numbers MC_UU_ 12023/21 and MC_ UU_ 12023/29.

CONFLICT OF INTEREST

James Carpenter reports grants from the Medical Research Council, UK, personal fees from Novartis, and a research grant from Astra Zeneca, during the conduct of the study. The other authors report no conflicts of interest.

ETHICS STATEMENT

This study is a re-analysis of a previously published trial-based economic evaluation and does not require any additional ethical approval.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Alexina J. Mason https://orcid.org/0000-0001-7319-4545
Richard Grieve https://orcid.org/0000-0001-8899-1301

REFERENCES

Baio, G. (2013). *Bayesian methods in health economics*. Chapman & Hall.

Baio, G. (2014). Bayesian models for cost-effectiveness analysis in the presence of structural zero costs. *Statistics in Medicine*, *33*(11), 1900–1913.

Baio, G. (2020). survHE: Survival analysis for health economic evaluation and cost-effectiveness modelling. *Journal of Statistical Software*.

Basu, A., & Manca, A. (2012). Regression estimators for generic health-related quality of life and quality-adjusted life years. *Medical Decision Making*, *32*(1), 56–69.

Daniels, M. J., & Hogan, J. W. (2008). Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis. In *Chapter 8: Models for handling nonignorable missingness* (pp. 165–215). Chapman & Hall.

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(9), 1–25.

Faria, R., Gomes, M., Epstein, D., & White, I. R. (2014). A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. *PharmacoEconomics*, *32*, 1157–1170.

Gabrio, A., Daniels, M. J., & Baio, G. (2020). A Bayesian parametric approach to handle missing longitudinal outcome data in trial-based health economic evaluations. *Journal of the Royal Statistical Society: Series A*, *183*(2), 607–629.

Gabrio, A., Mason, A. J., & Baio, G. (2017). Handling missing data in within-trial cost-effectiveness analysis: A review with future recommendations. *PharmacoEconomics Open*, *1*, 79–97.

Gabrio, A., Mason, A. J., & Baio, G. (2019). A full Bayesian model to handle structural ones and missingness in economic evaluations from individual-level data. *Statistics in Medicine*, *38*(8), 1399–1420.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Glick, H. A., Doshi, J. A., Sonnad, S. S., & Polsky, D. (2007). *Economic evaluation in clinical trials*. Oxford University Press.

Gomes, M., Radice, R., Brenes, J. C., & Marra, G. (2019). Copula selection models for non-Gaussian outcomes that are missing not at random. *Statistics in Medicine*, *38*(3), 480–496.

Grant, A. M., Boachie, C., Cotton, S. C., Faria, R., Bojke, L., Epstein, D. M., Ramsay, C. R., Corbacho, B., Sculpher, M., Krukowski, Z. H., Heading, R. C., & Campbell, M. K. (2013). Clinical and economic evaluation of laparoscopic surgery compared with medical management for gastro-oesophageal reflux disease: 5-year follow-up of multicentre randomised trial (the REFLUX trial). *Health Technology Assessment*, *17*(22), 1–167.

Grieve, R., Nixon, R., & Thompson, S. G. (2010). Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. *Medical Decision Making*, *30*, 163–175.

Guyot, P., Ades, A. E., Beasley, M., Lueza, B., Pignon, J. P., & Welton, N. J. (2017). Extrapolation of survival curves from cancer trials using external information. *Medical Decision Making*, *37*(4), 353–366. https://doi.org/10.1177/0272989X16670604

Hernandez-Alava, M., Wailoo, A. J., & Ara, R. (2012). Tails from the peak district: Adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. *Value in Health*, *15*, 550–561.

Jackson, C. H. (2008). Displaying uncertainty with shading. *The American Statistician*, *62*(4), 340–347.

Kharroubi, S. A., Edlin, R., Meads, D., Browne, C., Brown, J., & McCabe, C. (2015). Use of Bayesian Markov chain Monte Carlo methods to estimate EQ-5D utility scores from EORTC QLQ data in myeloma for use in cost-effectiveness analysis. *Medical Decision Making*, *35*(3), 351–360.

Kharroubi, S. A., Edlin, R., Meads, D., & McCabe, C. (2018). Bayesian statistical models to estimate EQ-5D utility scores from EORTC QLQ data in myeloma. *Pharmaceutical Statistics*, *17*(4), 358–371.

Kharroubi, S. A., O'Hagan, A., & Brazier, J. E. (2005). Estimating utilities from individual health preference data: A nonparametric Bayesian method. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *54*(5), 879–895.

Lambert, P. C., Billingham, L. J., Cooper, N. J., Sutton, A. J., & Abrams, K. R. (2008). Estimating the cost-effectiveness of an intervention in a clinical trial when partial cost information is available: A Bayesian approach. *Health Economics*, *17*, 67–81.

Leurent, B., Gomes, M., & Carpenter, J. (2018a). Missing data in trial-based cost-effectiveness analysis: An incomplete journey. *Health Economics*, *27*(6), 1024–1040.

Leurent, B., Gomes, M., Cro, S., Wiles, N., & Carpenter, J. (2020). Reference-based multiple imputation for missing data sensitivity analyses in trial-based cost-effectiveness analysis. *Health Economics*, *29*(2), 171–184.

Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R., & Carpenter, J. (2018b). Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: A tutorial. *PharmacoEconomics*, *36*(8), 889–901.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). Chapter 5: Prior distributions (pp. 81–102). Chapman & Hall.*The BUGS book: A practical introduction to Bayesian analysis*.

Mason, A., Richardson, S., Plewis, I., & Best, N. (2012). Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. *Journal of Official Statistics*, *28*(2), 279–302.

Mason, A. J., Gomes, M., Grieve, R., & Carpenter, J. (2018). A Bayesian framework for health economic evaluation in studies with missing data. *Health Economics*, *27*, 1670–1683. https://doi.org/10.1002/hec.3793

Mihaylova, B., Briggs, A., O'Hagan, A., & Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, *20*(8), 897–916.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiastis, A., & Verbeke, G. (Eds.). (2015). *Handbook of missing data methodology*. Chapman & Hall.

Nixon, R. M., & Thompson, S. G. (2005). Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics*, *14*, 1217–1229.

O'Hagan, A., & Stevens, J. W. (2001). A framework for cost-effectiveness analysis from clinical trial data. *Health Economics*, *10*, 303–315.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*. March 20-22, Vienna, Austria.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing

Rutherford, M. J., Lambert, P. C., Sweeting, M. J., Pennington, B., Crowther, M. J., Abrams, K. R., & Latimer, N. R. (2020). NICE DSU Technical Support Document 21: Flexible methods for survival analysis report BY the decision support unit 23 January 2020 SCHARR Univeristy of Sheffield.

Sanders, G. D., Neumann, P. J., Basu, A., Brock, D. W., Feeny, D., Krahn, M., Kuntz, K. M., Meltzer, D. O., Owens, D. K., Prosser, L. A., Salomon, J. A., Sculpher, M. J., Trikalinos, T. A., Russell, L. B., Siegel, J. E., & Ganiats, T. G. (2016). Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: Second panel on cost-effectiveness in health and medicine. *Journal of the American Medical Association*, *316*(10), 1093–1103.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B*, *64*(4), 583–639.

Thompson, S. G., & Nixon, R. M. (2005). How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Medical Decision Making*, *25*, 416–423.

Thorn, J. C., Davies, C. F., Brookes, S. T., Noble, S. M., Dritsak, M., Gray, E., Hughes, D. A., Mihaylova, B., Petrou, S., Ridyard, C., Sach, T., Wilson, E. C. F., Wordsworth, S., & Hollingworth, W. (2020). Content of health economics analysis plans (HEAPs) for trial-based economic evaluations: Expert Delphi Consensus Survey. *Value in Health*. https://doi.org/10.1016/j.jval.2020.10.002

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Mason, A. J., Gomes, M., Carpenter, J., & Grieve, R. (2021). Flexible Bayesian longitudinal models for cost-effectiveness analyses with informative missing data. *Health Economics*, *30*(12), 3138–3158. https://doi.org/10.1002/hec.4408

APPENDIX

A1 MODEL IMPLEMENTATION

We used the statistical software R (R Core Team, 2020) for pre-processing the trial data and post-processing the posterior samples, which were generated using the JAGS software (Plummer, 2003), called via the R package runjags (Denwood, 2016). All the models were run with two chains initialised using diffuse starting values to produce a sample of 10000 after convergence for posterior inference, providing an effective sample size of at least 3000 for the quantities of interest (thinning set to 10). Convergence is assumed if the potential scale reduction factor of the Gelman-Rubin statistic (Gelman & Rubin, 1992) is less than 1.05 for individual model parameters and a visual inspection of the trace plot for each parameter is satisfactory.

To implement the hurdle model for analysing HRQoL, we created a zero value indicator, h, set to 1 if HRQoL = 1 (i.e. HRQoL decrement is 0) and 0 otherwise. We explored different parameterisations of the gamma model for prior specification, and recommend using the shape and mean as this reduces the correlation between parameters compared to other options (e.g. using the mean and sd led to very high correlation between the mean intercept and sd). Also, in line with the usual recommendations for fitting Bayesian models, we scaled and centered all covariates, including binary variables. An upper limit of 1.6 was imposed on the gamma distributions, consistent with the range of legitimate values for EQ-5D.

A1.1 Selection of priors

We generally select minimally informative priors, the exception being the 'sensitivity type' parameters in the response missingness model that control the amount of departure from MAR (Section 2.5). Following the recommendations in Lunn et al. (2013) for logistic regression models, we place logistic(0, 1) priors on the intercept and normal(0, 1.65^2) priors on the other regression coefficients (normal parameterised in terms of the mean and variance). These generate approximately flat priors on the probability scale. For other regression models, the location and scale parameters are given normal(0, 10^2) priors and uniform(0, 100) priors respectively. As any correlation between the endpoints is expected to be negative, we restrict the prior on the ξ parameter (Equation (2)) to negative values, normal(0, 10^2)T(, 0).

A2 CEA OUTPUT EQUATIONS

Here, we provide the equations for the CEA outputs sub-model. We no longer suppress the treatment subscript, tr ($tr = 1$ denotes MM; $tr = 2$ denotes LS), and to simplify we denote the baseline covariates for patient i as vector $\boldsymbol{X}_i$.

Estimate HRQoL decrement assuming patients receive treatment 1 (MM) as follows:

$$\begin{aligned} h_{1,it} &\sim \text{Bernoulli}(p_{1,it}) \\ \text{logit}(p_{1,it}) &= \gamma_{t,tr=1} + \boldsymbol{\omega}_{tr=1}^{T}\boldsymbol{X}_i \\ \mu.q_{1,it} &= \exp(\alpha_{t,tr=1} + \theta_{1,i} + \beta_{tr=1}^{T}\boldsymbol{X}_i) \times (1 - h_{1,it}) \\ \theta_{1,i} &\sim \text{Normal}(\theta.\mu_{tr=1}, \theta.\sigma_{tr=1}^{2}). \end{aligned}$$

Estimate the HRQoL decrement assuming patients receive treatment 2 (LS) analogously.

Predict HRQoL decrement, assuming cross-over patients always receive the treatment they received during the trial

$$pred.q_{1,it} = (\mu.q_{1,it} \times (1 - MMxover_i)) + (\mu.q_{2,it} \times MMxover_i)$$
$$pred.q_{2,it} = (\mu.q_{2,it} \times (1 - LSxover_i)) + (\mu.q_{1,it} \times LSxover_i)$$

where *MMxover* is a binary indicator variable set to 1 if the patient was randomised to MM but received LS; and 0 otherwise. *LSxover* is defined analogously.

Estimate HRQoL differences: $diff.q_i = pred.q_{1,i} - pred.q_{2,i}$.

Estimate 1-year QALY differences

$$diff.Q_{iy} = \begin{cases} (0.5 \times diff.q_{iy}) + (0.375 \times diff.q_{i(y+1)}) & y = 1 \\ 0.5 \times (diff.q_{iy} + diff.q_{i(y+1)})/discount^{y-1} & y > 1. \end{cases}$$

Estimate 5-year QALY differences: $diff.Qtot_i = \sum_y diff.Q_{iy}$.

Estimate 5-year QALY increment: $Qinc = \frac{1}{N}\sum_i diff.Qtot_i$, where N is the total number of patients in the trial.

Estimate costs assuming patients receive treatment 1 (MM): $\mu.c_{1,i} = \exp(\zeta_{0,tr=1} + \boldsymbol{\zeta}^T_{tr=1}\boldsymbol{X}_i)$, and estimate costs assuming patients receive treatment 2 (LS) analogously.

Predict costs, assuming cross-over patients always receive the treatment they received during the trial

$$pred.c_{1,i} = (\mu.c_{1,i} \times (1 - MMxover_i)) + (\mu.c_{2,i} \times MMxover_i)$$
$$pred.c_{2,i} = (\mu.c_{2,i} \times (1 - LSxover_i)) + (\mu.c_{1,i} \times LSxover_i).$$

Estimate cost differences

$$diff.C_i = pred.c_{1,i} - pred.c_{2,i}$$
$$Cinc = \frac{1}{N}\sum_i diff.C_i.$$

Estimate incremental net benefits (INB), valuing quality-adjusted life year gains at 20,000GBP per quality-adjusted life year

$$INB = (20000 \times Qinc) - Cinc.$$

A3 JAGS JOINT MODEL CODE

```
# JAGS joint model of CEA for the REFLUX trial - QALYs allowed to be MNAR (selection model)
# analysis model for QALYs and Costs, 2 arms separately parameterised
#     HRQoL: hurdle at 0, with Gamma for non-zeros (priors on shape (shape.q) and mean (mu.q) coefficients)
#     Costs: gamma
#     HRQoL part includes fixed time effects, baseline HRQoL decrement (bq) and covariates (X),
#        and individual random effects in non-zeros model
#     Cost part includes covariates and is conditional on QALYs
#         and incorporates information from partially observed costs by imposing a lower bound
# covariate imputation model for baseline HRQoL (bq)
# response missingness model distinguishes 2 types of missingness using multinomial logistic model
#    all parameters vary by treatment arm
#    includes time fixed effects, covariates (X), last HRQoL observation (previous.q)
#        and change from last HRQoL observation (change.q)
# q = HRQoL decrements (1-HRQoL)
# h = 1 if q = 0; 0 if q > 0 (zero value indicator)
# cost = total costs in £1000s over 5 years
# treatment arm is indexed by tr in this model (1 = MM and 2 = LS)
# MMxover = 1 if patient randomised to MM but receives LS; 0 otherwise
# LSxover = 1 if patient randomised to LS but receives MM; 0 otherwise
```

```
# prepare data for multinomial logistic response missingness sub-model
data {
   for (i in 1:Np) { # loop through individuals
      for (t in 1:Nt) { # loop through timepoints provided patients have not already dropped out
         for (r in 1:3) {count[i,t,r] <- equals(mind[i,t],r)} # set up multinomial count
      }
   }
}

model{

   # ***** marginal model for health outcome *****

   # specify marginal hurdle sub-model for HRQoL decrements
   for (t in 1:Nt) { # 6 timepoints
      for (i in 1:Np) { # Np individuals
         h[i,t] ~ dbern(p[i,t]) # 1 indicates zeros model
         logit(p[i,t]) <- gamma[t,tr[i]] + omega0[tr[i]]*bqC[i] + inprod(omega[1:Nx,tr[i]],X[i,1:Nx])
         d[i,t] <- h[i,t]+1 # model index (1 = MM and 2 = LS)
         q[i,t] ~ dgamma(shape.q[d[i,t],tr[i]],rate.q[d[i,t],i,t])T(,1.6)
         pred.h[i,t] ~ dbern(p[i,t])
         pred.q[i,t] <- (1-pred.h[i,t])*mu.q[i,t] # prediction is 0 if h=1

         # non-zeros model
         log(mu.q[i,t]) <- alpha[t,tr[i]] + theta[i] + beta0[tr[i]]*bqC[i] + inprod(beta[1:Nx,tr[i]],X[i,])
         rate.q[1,i,t] <- shape.q[1,tr[i]]/mu.q[i,t]

         # zeros model - not used in QALY increment calculation
         rate.q[2,i,t] <- shape.q[2,tr[i]]/mu.q0

         # calculate residuals
         resid[i,t] <- q[i,t] - pred.q[i,t]

         # predict HRQoL assuming all participants have treatment 1 (MM)
         h1[i,t] ~ dbern(p1[i,t])
         logit(p1[i,t]) <- gamma[t,1] + omega0[1]*bqC[i] + inprod(omega[1:Nx,1],X[i,1:Nx])
         mu.q1[i,t] <- exp(alpha[t,1]+theta1[i]+beta0[1]*bqC[i]+inprod(beta[1:Nx,1],X[i,1:Nx])) * (1-h1[i,t])# 0 if h0=1
         # prediction assuming cross-overs receive LS
         pred.q1[i,t] <- (mu.q1[i,t] * (1-MMxover[i])) + (mu.q2[i,t] * MMxover[i])

         # predict HRQoL assuming all participants have treatment 2 (LS)
         h2[i,t] ~ dbern(p2[i,t])
         logit(p2[i,t]) <- gamma[t,2] + omega0[2]*bqC[i] + inprod(omega[1:Nx,2],X[i,1:Nx])
         mu.q2[i,t] <- exp(alpha[t,2]+theta2[i]+beta0[2]*bqC[i]+inprod(beta[1:Nx,2],X[i,1:Nx])) * (1-h2[i,t]) # 0 if h0=1
         # prediction assuming cross-overs receive MM
         pred.q2[i,t] <- (mu.q2[i,t] * (1-LSxover[i])) + (mu.q1[i,t] * LSxover[i])

         # calculate HRQoL differences
         diff.q[i,t] <- pred.q1[i,t] - pred.q2[i,t] # difference is LS-MM HRQoL (switch from HRQoL decrement)
      }
   }

   for (i in 1:Np) { # individual random effects for HRQoL marginal model
```

```
   theta[i] ~ dnorm(theta.mu[tr[i]],theta.tau[tr[i]])
   theta1[i] ~ dnorm(theta.mu[1],theta.tau[1]) # random effects for treatment 1 predictions
   theta2[i] ~ dnorm(theta.mu[2],theta.tau[2]) # random effects for treatment 2 predictions
}

# prior distributions for HRQoL marginal sub-model
for (a in 1:2) { # 2 treatment arms
   for (t in 1:Nt) {gamma[t,a] ~ dlogis(0,1)} # time fixed effects for hurdle
   omega0[a] ~ dnorm(0,0.368)
   for (i in 1:Nx) {omega[i,a] ~ dnorm(0,0.368)}
   alpha[1,a] <- 0
   for (t in 2:Nt) {alpha[t,a] ~ dnorm(0,0.01)} # time fixed effects for non-zeros model
   beta0[a] ~ dnorm(0,0.01)
   for (i in 1:Nx) {beta[i,a] ~ dnorm(0,0.01)}
   shape.q[1,a] ~ dunif(0,100)

   theta.mu[a] ~ dnorm(0,0.01) # prior on random effects mean
   theta.sigma[a] ~ dunif(0,100) # prior on random effects sd

   # node transformations
   theta.sigma2[a] <- pow(theta.sigma[a],2)
   theta.tau[a] <- 1/theta.sigma2[a]
}

# set mean and sd of zeros model to induce a spike close to 0
mu.q0 <- 0.0001
for (a in 1:2) {shape.q[2,a] <- 0.0001}

# ***** conditional model for cost outcome *****

# specify conditional gamma sub-model for costs
for (i in 1:Np) { # Np individuals
   # switch from HRQoL decrements to HRQoL to calculate QALYs
   Qtot[i] <- (0.5*(1-q[i,1])) + (0.875*(1-q[i,2])) + sum((1-q[i,3:5])) + (0.5*(1-q[i,6]))
   Qmu[i] <- (0.5*(1-pred.q[i,1])) + (0.875*(1-pred.q[i,2])) + sum((1-pred.q[i,3:5])) + (0.5*(1-pred.q[i,6]))

   # model costs conditional on QALYS
   cost[i] ~ dgamma(shape.c[tr[i]],rate.c[i])T(lower[i],)
   log(mu.c[i]) <- zeta0[tr[i]] + inprod(zeta[1:Nx,tr[i]],X[i,]) + xi[tr[i]]*(Qtot[i]-Qmu[i])
   rate.c[i] <- shape.c[tr[i]]/mu.c[i]

   # predict cost assuming all participants receive treatment 0
   mu.c1[i] <- exp(zeta0[1] + inprod(zeta[1:Nx,1],X[i,1:Nx]))
   # prediction assuming cross-overs receive LS
   pred.c1[i] <- (mu.c1[i] * (1-MMxover[i])) + (mu.c2[i] * MMxover[i])

   # predict cost assuming all participants receive treatment 1
   mu.c2[i] <- exp(zeta0[2] + inprod(zeta[1:Nx,2],X[i,1:Nx]))
   # prediction assuming cross-overs receive MM
   pred.c2[i] <- (mu.c2[i] * (1-LSxover[i])) + (mu.c1[i] * LSxover[i])

   # calculate cost differences
```

```
   Ctot.diff[i] <- pred.c2[i] - pred.c1[i]
}

# prior distributions for cost conditional sub-model
for (a in 1:2) { # 2 treatment arms
   zeta0[a] ~ dnorm(0,0.01)
   for (i in 1:Nx) {zeta[i,a] ~ dnorm(0,0.01)}
   xi[a] ~ dnorm(0,0.01)T(,0) # any correlation expected to be negative
   shape.c[a] ~ dunif(0,100)
}

# ***** covariate imputation model *****

# specify covariate imputation model for baseline HRQoL decrement
for (i in 1:Np) {
   bq[i] ~ dgamma(shape.bq,rate.bq)T(,1.2)
   # center and standardise for HRQoL marginal sub-model
   bqC[i] <- (bq[i]-mean.bq)/sd.bq
}

# prior distributions for covariate imputation model
shape.bq ~ dunif(0,100)
mu.bq ~ dunif(0,1.2)
rate.bq <- shape.bq/mu.bq

# ***** health outcome missingness model *****

# specify response missingness model for HRQoL
for (i in 1:Np) { # loop through individuals
   previous.q[i,1] <- bq[i]
   for (t in 2:Nt) {previous.q[i,t] <- q[i,t-1]}
   for (t in 1:Nt) { # loop through all timepoints
      count[i,t,1:3] ~ dmulti(m[i,t,1:3],1)
      change.q[i,t] <- q[i,t]-previous.q[i,t]
      for (r in 1:3) {
         m[i,t,r] <- phi[i,t,r]/sum(phi[i,t,])
         log(phi[i,t,r]) <- kappa0[r,t,tr[i]] + inprod(kappa[r,1:Nx,tr[i]],X[i,1:Nx])
            + kappa[r,4,tr[i]]*(previous.q[i,t]-mean.q)/sd.q + lambda[r,tr[i]]*change.q[i,t]
      }
   }
}

# prior distributions for response missingness model
for (a in 1:2) { # 2 treatment arms
   for (t in 1:Nt) {kappa0[1,t,a] <- 0}
   for (i in 1:4) {kappa[1,i,a] <- 0}
   for (r in 2:3) {
      for (t in 1:Nt) {kappa0[r,t,a] ~ dlogis(0,1)}
      for (i in 1:4) {kappa[r,i,a] ~ dnorm(0,0.01)}
   }
}

# ***** calculation of CEA outputs model *****
```

```
   # calculate QALY differences over 5 year period
   for (i in 1:Np) { # Np individuals
      Qaly1[i,1] <- (0.125*(1-bq[i])) + (0.5*(1-pred.q1[i,1])) + (0.375*(1-pred.q1[i,2])) # QALY for MM in year 1
      Qaly2[i,1] <- (0.125*(1-bq[i])) + (0.5*(1-pred.q2[i,1])) + (0.375*(1-pred.q2[i,2])) # QALY for LS in year 1
      Q.diff[i,1] <- (0.5*diff.q[i,1]) + (0.375*diff.q[i,2]) # QALY difference in year 1
      for (y in 2:5) { # years 2 to 5, applying discount
         Qaly1[i,y] <- 0.5 *(2-pred.q1[i,y]-pred.q1[i,y+1]) / pow(disc,y-1) # QALY for MM
         Qaly2[i,y] <- 0.5 *(2-pred.q2[i,y]-pred.q2[i,y+1]) / pow(disc,y-1) # QALY for LS
         Q.diff[i,y] <- 0.5 *(diff.q[i,y]+diff.q[i,y+1]) / pow(disc,y-1) # QALY difference
      }
      Qtot1[i] <- sum(Qaly1[i,]) # 5-year QALYs for MM
      Qtot2[i] <- sum(Qaly2[i,]) # 5-year QALYs for LS
      Qtot.diff[i] <- sum(Q.diff[i,]) # 5-year QALY difference
   }

   # calculate QALY increment using recycled predictions
   AveQ[1] <- mean(Qtot1[])
   AveQ[2] <- mean(Qtot2[])
   Qinc <- mean(Qtot.diff[]) # 5-year QALY increment
   p.Qinc <- step(Qinc) # probability favours LS
   for (y in 1:5) { # calculate QALY increment by year
      AveQ.1yr[y,1] <- mean(Qaly1[,y])
      AveQ.1yr[y,2] <- mean(Qaly2[,y])
      Q1yr.inc[y] <- mean(Q.diff[,y]) # 1-year QALY increment
      p.Q1yr[y] <- step(Q1yr.inc[y]) # probability favours LS
   }

   # calculate cost increment using recycled predictions
   AveC[1] <- mean(pred.c1[])*1000
   AveC[2] <- mean(pred.c2[])*1000
   Cinc <- mean(Ctot.diff[])*1000 # 5-year cost increment
   p.Cinc <- 1-step(Cinc) # probability favours LS (Cinc positive)

   # calculate incremental net benefits (INB)
   for (j in 1:M) { # values of efficacy (QALY) gains
         inb[j] <- (threshold[j]*Qinc)-Cinc
         p.ce[j] <- step(inb[j]) # probability favours LS (INB positive)
   }
}
```