BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Dirichlet process mixture models to impute missing predictor data in counterfactual prediction models: an application to predict optimal type 2 diabetes therapy

Pedro Cardoso[1], John M. Dennis[1], Jack Bowden[1], Beverley M. Shields[1], Trevelyan J. McKinley[1*] and the MASTERMIND Consortium

Abstract

Background The handling of missing data is a challenge for inference and regression modelling. A particular challenge is dealing with missing predictor information, particularly when trying to build and make predictions from models for use in clinical practice.

Methods We utilise a flexible Bayesian approach for handling missing predictor information in regression models. This provides practitioners with full posterior predictive distributions for both the missing predictor information (conditional on the observed predictors) and the outcome-of-interest. We apply this approach to a previously proposed counterfactual treatment selection model for type 2 diabetes second-line therapies. Our approach combines a regression model and a Dirichlet process mixture model (DPMM), where the former defines the treatment selection model, and the latter provides a flexible way to model the joint distribution of the predictors.

Results We show that DPMMs can model complex relationships between predictor variables and can provide powerful means of fitting models to incomplete data (under missing-completely-at-random and missing-at-random assumptions). This framework ensures that the posterior distribution for the parameters and the conditional average treatment effect estimates automatically reflect the additional uncertainties associated with missing data due to the hierarchical model structure. We also demonstrate that in the presence of multiple missing predictors, the DPMM model can be used to explore which variable(s), if collected, could provide the most additional information about the likely outcome.

Conclusions When developing clinical prediction models, DPMMs offer a flexible way to model complex covariate structures and handle missing predictor information. DPMM-based counterfactual prediction models can also provide additional information to support clinical decision-making, including allowing predictions with appropriate uncertainty to be made for individuals with incomplete predictor data.

A list of authors (the MASTERMIND Consortium) and their affiliations appears at the end of the paper.

*Correspondence:
Trevelyan J. McKinley
t.mckinley@exeter.ac.uk
Full list of author information is available at the end of the article

Keywords Dirichlet process mixture model, Treatment selection model, Precision medicine, Type 2 diabetes, Bayesian modelling

Background

Prediction models are being increasingly deployed to support clinical decision making in healthcare. These models typically take a set of predictor (explanatory) variables—for example, a set of routinely collected clinical features—and then use these to directly predict a key outcome-of-interest. Counterfactual prediction models provide a recent extension to standard prediction models by supporting estimation of differential effects of treatments conditional on predictor variables. Models predicting such differential treatment effects have the potential to inform precision medicine approaches aiming to target specific treatment(s) to individual patient most likely to benefit based on their characteristics. Such models are often termed 'treatment selection models'. A major challenge when building and implementing both types of clinical prediction models in practice is dealing with missing predictor data. This is particularly pertinent if a given model includes predictor variables that are informative about the outcome, but are not routinely collected, such as diagnostic test information. In this case it is not possible to produce predictions unless the model has a way to impute the missing information, thus severely hampering the utility of such models when deployed in clinical practice.

In the context of counterfactual prediction models, dealing robustly with missing data is one of the key priority research areas highlighted in the recent Predictive Approaches to Treatment effect Heterogeneity (PATH) statement, a document outlining principles, criteria and key considerations for predictive approaches of heterogeneity of treatment effects [1]. In particular, this manuscript aims to contribute to the requirement to: "Determine methods to permit models predicting treatment effect to cope with missing data in clinical practice" [1]. In this study, we extend an exemplar recently published counterfactual treatment selection model for type 2 diabetes (T2D) that predicts optimal treatment for individual patients based on routinely collected clinical features [2]. We build on this previous work (which was based on a complete-case analysis) to explore the potential for using cutting-edge Bayesian approaches to handle missing data during both model development and model deployment.

There are three types of missing data that occur during data collection: data can be missing completely-at-random (MCAR) when the missing values can be thought of as a random sub-sample of the actual values; missing-at-random (MAR) when the pattern of missingness is predictable from other variables in the dataset; and missing-not-at-random (MNAR) when the missingness is non-random, and it is not predictable from other variables in the dataset [3]. Due to difficulties modelling missing data, many prediction models use only complete-case information, which can lead to biases unless the data are MCAR [4]. The complete-case analysis also throws away potentially informative data from individuals with incomplete records. Since clinical data is often incomplete, then without a way to handle missing predictors, it would also not be possible to predict for any new patient with incomplete predictor information unless there was a way to robustly impute the missing data.

A common approach for dealing with missing predictors, under the assumption of MAR, involves the use of multiple imputation (MI), commonly through chained equations (MICE) [3, 5]. Whilst powerful and straightforward to implement, MICE has some limitations. Unless the chained equations have a carefully chosen hierarchical dependence structure, these do not always provide a probabilistically consistent joint probability model for the predictors [6]. Another limitation of multiple imputation involves the propagation of uncertainties to the parameters and predictions, since this only occurs through the standard errors and therefore does not provide the richness of information afforded by full Bayesian posterior (predictive) distributions.

Furthermore there are various challenges with the deployment of prediction models using MI in clinical practice [7]. For example, it is often recommended that the response variable should be included in the MI model to improve the accuracy of the fitted model [8]; however, by definition the response variable will not be observed if using the model to predict the outcome for a new individual. As such, to produce predictions for new individuals with missing data, then MI would have to be conducted without using the response variable, thus potentially reducing the accuracy of the predictions [8]. For more discussion of these challenges, please see [7].

Bayesian approaches offer principled ways to deal with the issues described above. The implementation of a hierarchical Bayesian prediction model, in which the response model is dependent on a flexible joint probability model for the predictor variables, facilitates both inference and prediction in the presence of missing data

[9]. Firstly, this defines a joint model for the response and predictor distributions, meaning that information on the response variable is naturally incorporated into posterior estimates of missing predictors during model fitting. Furthermore, one can estimate full posterior predictive distributions for new individuals empirically, using posterior samples obtained via e.g MCMC methods. This has a useful side-effect in that for a given set of posterior samples, it is possible to generate samples from a corresponding predictive distribution for a new individual without requiring access to the original dataset. This is in contrast to models affording an analytical solution, such as classical linear regression, where confidence and prediction intervals are typically functions of the original data. This makes general deployment of such models easier in the case where the original training data is under confidentiality restrictions with restricted access.

To this end, Dirichlet process mixture models (DPMMs) [10] provide a powerful way to model the joint (complex) distribution of a set of predictor variables. They do this by modelling complex non-standard distributions through a mixture of simpler, more tractable distributions (see Table 1). This approach can readily capture non-standard features such as non-linearities and heteroscedasticity across different dimensions. Furthermore, it can handle mixtures of both numerical and categorical predictors [11–15]. This allows for fully probabilistic inference under the MCAR and MAR assumptions, contributing towards addressing some of the points raised in the PATH statement highlighted above.

For an extensive exploration of Bayesian non-parametric methods and missing data, see [20]. DPMMs have also been established as a clustering mechanism [21], used in multiple imputation [14, 15, 22–26] and used as a joint model in a specific type of hierarchical regression model known as profile regression [11, 12]. In *profile regression*, a DPMM is fitted to a set of predictor variables providing a probabilistic mapping of each individual observation to different component clusters within the DPMM. The cluster labels are then used as predictor variables in a hierarchical regression model targeting the response. Hence, profile regression ultimately captures the relationship between covariate profiles (as modelled through the cluster structure of the DPMM) and the response variable-of-interest [11]. In contrast, in this manuscript, we use a DPMM to explicitly model the joint distribution of the predictors (given the observed data), which can then impute missing data directly and feed these into the outcome model-of-interest. This is similar to the approach used in [27], who develop a flexible hierarchical framework that can be applied to treatment selection models that uses an enriched Dirichlet process [28, 29] as a means of capturing the complex joint distribution

Table 1 Dirichlet process mixture models (DPMMs)

A DPMM is commonly used as a prior distribution over the components of a (possibly multivariate) mixture model of unknown complexity [16]. A DPMM consists of a theoretically infinite number of components, where each component is parameterised by a specific functional form with component-specific parameters. Instead of fitting an infinite number of components, we use a truncated DPMM with a maximum number of components K, based on the assumption the optimal number of components is lower than the set limit [17]. In turn, the computational demands for fitting the model are reduced. The value of K should increase as the complexity of the data increases, but its suitability can be checked post model fit.

The DPMM thus defines a weighted sum of K component densities [18]. The component densities are restricted to particular parametric classes of densities that are assumed to be appropriate for the data at hand. We define $f_k(\mathbf{X} \mid \boldsymbol{\Theta}_k)$ as the k th component density, with $\boldsymbol{\Theta}_k$ representing the component parameters. A K component mixture density is defined as:

$$f\left(\mathbf{X} \mid \boldsymbol{\Theta}\right) = \sum_{k=1}^{K} p_k f_k\left(\mathbf{X} \mid \boldsymbol{\Theta}_k\right),$$

where p_k are component-specific weights such that $\sum_{k=1}^{K} p_k = 1$[19].

For J_C continuous predictors, we use mixtures of multivariate Gaussian distributions with J_C dimensions, the cluster-specific parameters for component k ($k = 1, \ldots, K$) are given by $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k$ is a J_C-vector of means and $\boldsymbol{\Sigma}_k$ is a $(J_C \times J_C)$ covariance matrix. For J_D categorical predictors, we use mixtures of categorical probability mass functions, where the number of categories for a covariate j ($j = 1, \ldots, J_D$) is K_j, the component-specific parameters are the probabilities of belonging to each category, given by $\boldsymbol{\phi}_k = (\boldsymbol{\phi}_{k1}, \boldsymbol{\phi}_{k2}, ..., \boldsymbol{\phi}_{kJ_D})$ with $\boldsymbol{\phi}_{kj} = (\phi_{kj1}, \phi_{kj2}, \ldots, \phi_{kjK_j})$ and $\sum_{l=1}^{K_j} \phi_{kjl} = 1$. The model in this paper is given as a mixture of continuous and categorical variables, since $\mathbf{X}_i = \left(\mathbf{X}_i^C, \mathbf{X}_i^D, X_i^T\right)$, with $\mathbf{X}_i^C$ representing the continuous predictors, $\mathbf{X}_i^D$ corresponding to the categorical predictors and $\mathbf{X}_i^T$ representing the treatment taken. Hence in the notation of the Study overview section, $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K, p_1, \ldots, p_K)$, where the component-specific parameters are given by $\boldsymbol{\Theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\phi}_k)$.

A latent variable, $Z_i = 1, \ldots, K$, is used to assign individual data points to different components of the mixture model, and we assume independence between continuous and categorical components conditional on the cluster allocations [11, 12]. Thus the probability density for individual i, given Z_i is:

$$f(\mathbf{X}_i \mid Z_i, \boldsymbol{\Theta}_{Z_i}) = f_{Z_i}\left(\mathbf{X}_i \mid \boldsymbol{\Theta}_{Z_i}\right) = f(\mathbf{X}_i^C \mid \boldsymbol{\mu}_{Z_i}, \boldsymbol{\Sigma}_{Z_i}) f(\mathbf{X}_i^D \mid \boldsymbol{\phi}_{Z_i}) \qquad (1)$$

More details on the component densities, prior distributions, and how to sample from the DPMM are given in the Supplementary Materials.

of the predictors *and* the response, providing a means of imputing missing data under a MCAR/MAR assumption. These methods are highly flexible, and can generate estimates of the response regression function indirectly from the joint predictive density (an idea originating in [30]; see e.g. [29] and references therein for a more detailed discussion). However, they are technically challenging to implement in practice without bespoke code.

The original motivation for the work in this paper was to augment currently developed and validated models to enable prediction for new patients with incomplete predictor data, and thus facilitate their deployment in routine clinical practice. Furthermore we wanted to be able to develop these models in general-purpose Bayesian analysis software, to make these ideas more accessible to a wider audience and to make them applicable to a wide-class of pre-existing models. As such we decided to use a simpler approach where we use the DPMM to model the joint distribution of the predictor variables as a means of imputing missing data, and then use the aforementioned spline-based treatment selection model in T2D [2] as our outcome model-of-interest, which combines routine clinical features of each patient to predict future blood glucose (HbA1c) levels under two commonly prescribed treatments, SGLT2i and DPP4i. We will henceforth refer to this model as the SGLT2i-DPP4i treatment selection model. The analysis identified marked heterogeneity in HbA1c outcome that is predicted by baseline patient characteristics: age-at-treatment, baseline HbA1c, body mass index (BMI), estimated glomerular filtration rate (eGFR), alanine aminotransferase (ALT) measurements and previous treatment history. These predictions can determine which treatment is likely to be optimal for an individual patient based on their characteristics. We carefully validated the model in hold-out data, and well as the development data.

Furthermore, we show how the model can also be used to identify how predictions of the response or treatment selection outcomes might change if any of the missing predictor variable(s) were available, and if so, which ones (if any) would be the most useful to collect. We note that this latter feature could be particularly useful for modelling non-routinely collected diagnostic test information, such that one could examine how predictions might change if the test were to be conducted, and hence could be used to inform whether it is worthwhile to perform the test or not. We also provide code that enables this approach to be implemented in the general-purpose Bayesian modelling package NIMBLE, within the R statistical language (which by default does not currently allow for the conditional sampling of missing dimensions of multivariate nodes, which is required for this approach to work).

Methods

Study overview

The focus here is on enhancing the previously described SGLT2i–DPP4i penalised linear regression model designed to predict likely glucose-lowering response (HbA1c) in people with T2D initiating SGLT2i or DPP4i therapy [2]. For an individual patient, the difference in the predicted outcome for each treatment is then used to infer which is likely to be optimal. Such models are labelled treatment selection models [31].

The model uses interaction effects between the treatment and the other predictors to allow for differential responses to treatment given a set of predictors. For a new individual with a specific set of characteristics, mean predictions of the HbA1c change at 6/12 months post-treatment can be made under both treatments and then the difference can be derived. We note that there is considerable residual uncertainty in the fitted models, and as such [2, 31] only use point predictions when making comparisons. With this in mind, predictions should be interpreted as the expected difference for an individual with a specific set of predictors under the two treatments, commonly named the conditional average treatment effect (CATE). In the Bayesian model below, we follow the same principle, where posterior predictive distributions are for the expected treatment effect for an individual with a given set of characteristics.

In the model description below, we use $\mathbf{Y}$ to denote the outcome variable and $\mathbf{X}$ to denote the set of predictor variables. Hence, $f(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\psi})$ corresponds to the likelihood function from the regression model of Dennis et al. [2], with parameters $\boldsymbol{\psi}$; $f(\mathbf{X} \mid \boldsymbol{\Theta})$ corresponds to the joint probability density function for the DPMM, given parameters $\boldsymbol{\Theta}$ (see Table 1); and $f(\boldsymbol{\psi}, \boldsymbol{\Theta})$ corresponds to the prior distribution for the parameters, then our Bayesian hierarchical model has posterior distribution:

$$f(\boldsymbol{\psi}, \boldsymbol{\Theta}, \mathbf{X}^m \mid \mathbf{X}^o, \mathbf{Y}) \propto f(\mathbf{Y} \mid \mathbf{X}^o, \mathbf{X}^m, \boldsymbol{\psi}) f(\mathbf{X}^o, \mathbf{X}^m \mid \boldsymbol{\Theta}) f(\boldsymbol{\psi}, \boldsymbol{\Theta}). \tag{2}$$

where $\mathbf{X}^o$ corresponds to all observed covariates, the full set of covariates for complete individuals and the subset of covariates that were observed for those with missingness, and $\mathbf{X}^m$ represents the unobserved covariates, but not the observed values for the incomplete patients (see the Supplementary Materials Section S1 for more details). If fitting to complete-case data only, then the posterior simplifies to Equation (S.3) in the Supplementary Materials. We refer the reader to [9] for a comprehensive reference for Bayesian hierarchical models.

We fit this model using Markov chain Monte Carlo (MCMC) with the package NIMBLE [32, 33] (version 0.12.2) in the software **R** [34] (version 4.0.3). We note that in the current version of NIMBLE, using the in-built samplers, it is not possible to update individual dimensions of a multivariate node. However, NIMBLE does allow users to specify their own custom MCMC samplers, and so in order to implement model (2) we wrote our own custom samplers, and provide complete code for model fitting and prediction at https://github.com/PM-Cardoso/DPMM-tsm. We cannot share the original data due to confidentiality constraints, but we provide an example of using the code to fit a model to a smaller synthetic dataset.

Data/Study population

The analysis in this paper is performed with anonymised routinely collected, population-representative UK primary care electronic healthcare records from Clinical Practice Research Datalink (CPRD) GOLD (July 2019 download) [35], selecting users of SGLT2i and DPP4i therapy after January 1st, 2013. We used the same dataset as previously described by Dennis et al. [2] in the initial model development study and extracted the same clinical feature predictor variables (age-at-treatment, baseline HbA1c, BMI, eGFR, ALT, the number of previously prescribed glucose-lowering therapies, the current therapy and its duration, the number of ongoing prescribed treatments) and HbA1c outcome. The dataset is split into two groups: the development group (60% of the dataset corresponding to 16,126 patients) and the validation group (40% of the dataset corresponding to 10,751 patients). Compared to the dataset used to fit the original model [2], there are an extra 2,057 patients in the development group and 1,375 patients in the validation group with incomplete predictor variables (Table 2). We fit to the development dataset and undertake both internal and external validation.

Prediction model for treatment selection based on HbA1c outcome

The model from Dennis et al. (2022) [2] assumes a continuous, normally distributed response variable Y_i for $i = 1, \ldots, n$, where N corresponds to the number of observations. The model is a standard linear regression model with both continuous and categorical predictors ($\mathbf{X}_i^C$ and $\mathbf{X}_i^D$, respectively). Restricted cubic splines are placed on the continuous predictors ($S(\mathbf{X}_i^C)$) [36], and interaction terms are included between the binary variable for a drug taken (X_i^T) and the other variables. Hence the structure of the treatment selection model is:

$$\begin{aligned} Y_i &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_i^T + \boldsymbol{\beta}_2 \mathbf{X}_i^D + \boldsymbol{\beta}_3 S\left(\mathbf{X}_i^C\right) + \boldsymbol{\beta}_4 X_i^T \mathbf{X}_i^D + \boldsymbol{\beta}_5 X_i^T S\left(\mathbf{X}_i^C\right) + \epsilon_i \\ \epsilon_i &\sim N\left(0, \sigma^2\right), \end{aligned} \tag{3}$$

where $\boldsymbol{\beta}_0$ is the intercept, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_5\right)$ is a vector of regression coefficients (where $\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_5$ are themselves vectors of regression coefficients corresponding to

Table 2 Summaries of differential predictors and biomarkers used in the model for the development and validation datasets. (SD) [percentage of missing values]

	Development Dataset		**Validation Dataset**	
Drug Taken	DPP4-inhibitor (n = 9,974)	SGLT2-inhibitor (n = 6,152)	DPP4-inhibitor (n = 6,650)	SGLT2-inhibitor (n = 4,101)
Age (years)	63.9 (10.8)	59.9 (9.1)	65.0 (10.7)	60.2 (9.3)
Number of Past Drugs				
2	3,884	1,167	2,556	731
3	3,653	1,693	2,457	1,115
4	972	1,569	683	1,045
5+	186	945	117	672
Number of Current Drugs				
0	523	149	309	93
1	5,078	2,191	3,424	1,418
2	3,000	2,449	1,993	1,643
3+	94	585	87	409
HbA1c (mmol/mol)	72.9 (13.5)	76.6 (14.2)	72.6 (13.2)	77.1 (14.1)
eGFR (mL/min/1.3m2)	83.1 (17.4) [0.2%]	88.8 (14.7) [0.3%]	84.9 (17.2) [0.2%]	88.6 (14.8) [0.4%]
ALT (IU/L) (logged)	3.3 (0.5) [10.0%]	3.4 (0.5) [10.4%]	3.3 (0.5) [10.1%]	3.4 (0.5) [10.8%]
BMI (kg/m2)	32.3 (6.4) [3.2%]	34.4 (6.5) [2.4%]	32.3 (6.4) [2.9%]	34.4 (6.6) [2.5%]
Outcome HbA1c (mmol/mol)	65.1 (16.0)	64.9 (14.2)	65.0 (16.2)	65.1 (14.6)
HbA1c_Month: Month of outcome HbA1c	9.2 (3.5)	9.0 (3.5)	9.2 (3.4)	9.0 (3.4)

the different discrete, spline and interactions variables as required), with σ the residual standard deviation. Posterior summaries for the parameters can be found in Supplementary Table S1 and Fig. S5.

Hence, in the notation of the Study overview section, Y_i and X_i corresponds to the response and covariates for the individual i, such that $\mathbf{Y} = (Y_1, \ldots, Y_n)$ for the outcome, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ corresponds to the predictor values used in the modelling, where each observation's covariates can be decomposed into $\mathbf{X}_i = \left(\mathbf{X}_i^D, \mathbf{X}_i^C, X_i^T\right)$ and $\boldsymbol{\psi} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}, \sigma)$. For a full discussion of the form of this model, and the interpretation of the parameters, please see Dennis et al. (2022) [2].

Model validation

To assess convergence and mixing of the MCMC chains, we perform a visual inspection of the trace plots for all parameters, as well as monitoring the Gelman-Rubin $\hat{R}$ values [37] (Supplementary Figs. S2–S3). The Gelman-Rubin $\hat{R}$ values for α, σ and regression parameters vary between 1 and 1.005. Together these diagnostics suggest that the chains converged and are mixing well. The model took 3 hours to fit with complete data and 24 hours to fit with incomplete data.

Since the application of DPMM in the model uses a truncated number of clusters, we also inspect the posterior mean number of individuals in components ranked by occupancy (Supplementary Fig. S2) alongside a posterior predictive plot of the fitted DPMM against the empirical distribution of the predictors from the development and validation data sets (Supplementary Figs. S4 and 1 respectively). Contrasting standard prediction models, the principal aspect of a treatment selection model is to accurately predict the optimal treatment instead of solely focusing on the outcome [38].

Validation of the individualised treatment effects of the model is carried out according to the guidelines provided by Dennis et al. [2]. The first method suggested is verifying the prediction performance by checking standardised residuals for predicted HbA1c—Supplementary Fig. S6. This shows that there is no major systematic bias in the mean function but that there is considerable residual variation as per the results of [2]. The second method is to plot a density graph for all patients regarding the difference between each therapy's treatment effects (Supplementary Fig. S7) that can be compared to [2]. This shows the predicted number of patients benefiting from DPP4i or SGLT2i across the population and shows good agreement with [2].

Results

Fit and validation of the DPMM

To visualise the fit of the DPMM, we take the bivariate marginal posterior predictive distributions for a combination of variables versus the empirical equivalent (Fig. 1). The combination of density plots, density map plots and bar plots demonstrate that the fitted DPMM provides a good representation of both the development (Supplementary Fig. S4) and external validation (Fig. 1) datasets. We can see that the DPMM is very flexible and captures non-linearities, heteroscedasticity, correlation structures and mixtures of continuous and categorical variables.

Probabilistic predictions using the Bayesian treatment selection model

The Bayesian model produces full posterior predictive distributions that marginalise the uncertainty in the parameters and any missing predictor variables. This means we can make probabilistic statements about the parameter values and predicted outcomes based on the observed data and the choice of prior distributions. This is in contrast to frequentist approaches, where uncertainty statements relate to the expected long-term behaviour of an estimator if the same study were to be repeated *ad infinitum* [9]. In Fig. 2 (A), we showcase the posterior predictive distributions of average therapy response for three synthetic but representative patients (A, B and C) under the two competing treatments. An alternative way to visualise this information is to generate a posterior predictive distribution for the conditional average treatment effect (CATE) estimates (Fig. 2 (B)). Posterior samples for the CATE estimates can be calculated by subtracting the mean predicted treatment response of DPP4i from the mean predicted treatment response of SGLT2i (given the predictors) for each iteration of the fitted model. In this case, the proportion of the distribution to the left of the zero line corresponds to an estimate of the probability that SGLT2i is associated with a better average treatment response, and conversely, the proportion of the distribution to the right of the zero line is an estimate of the probability that DPP4i is associated with a better average treatment response. Hence the model predicts a 98% probability of SGLT2i performing better than DPP4i on average for patient A. For patient B, SGLT2i has a 58% probability of performing better on average than DPP4i, and for patient C the model predicts DPP4i as the optimal treatment on average with a >99% probability. With this approach, it is evident that there is a high probability of patient A having a 1–3 mmol/mol greater HbA1c reduction on average with SGLT2i instead

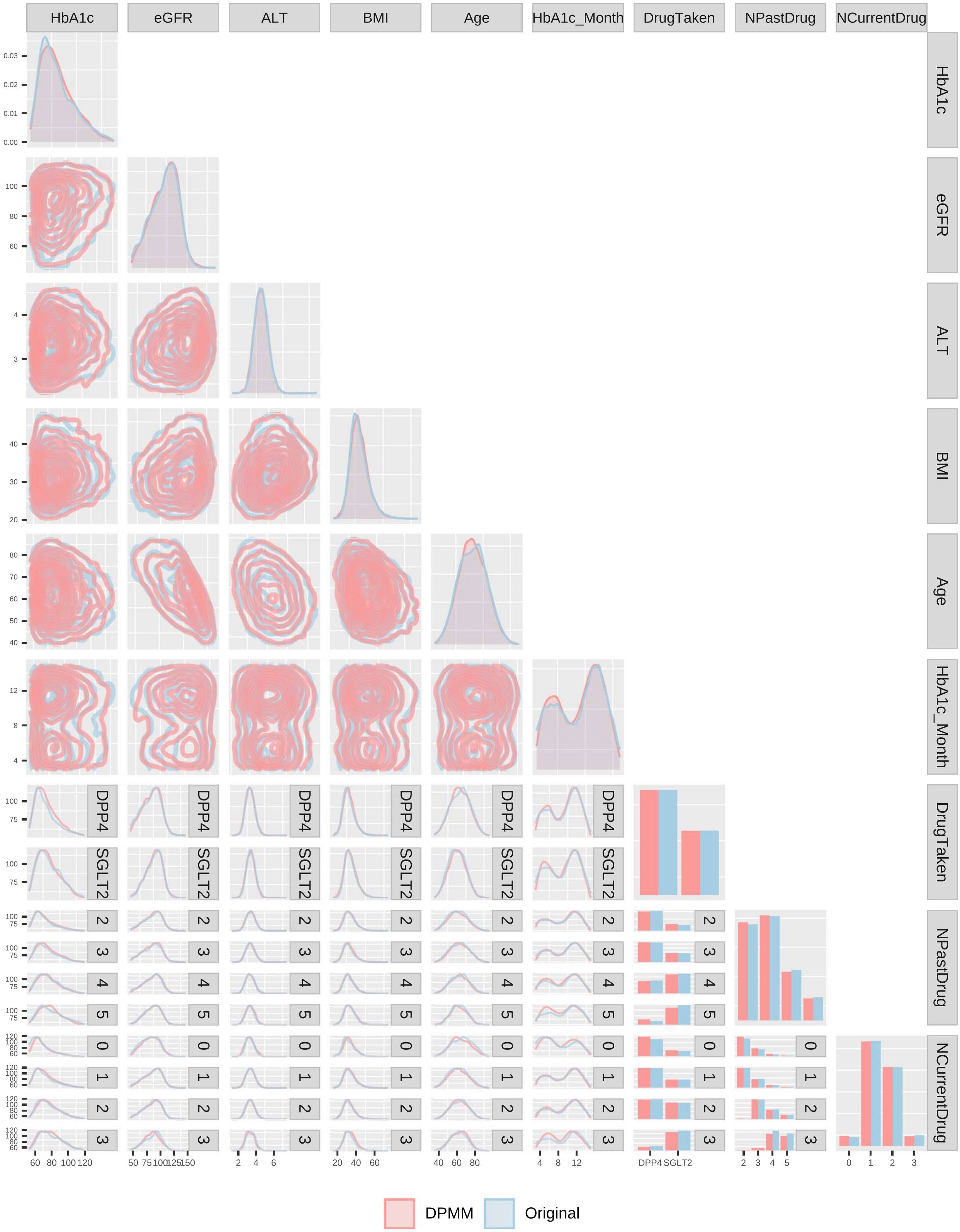

Fig. 1 Generalised pairs plot of predictor variables for an independent validation dataset against an equal number of random samples from the Dirichlet process mixture model (DPMM). The DPMM provides an exceptional representation of the validation dataset

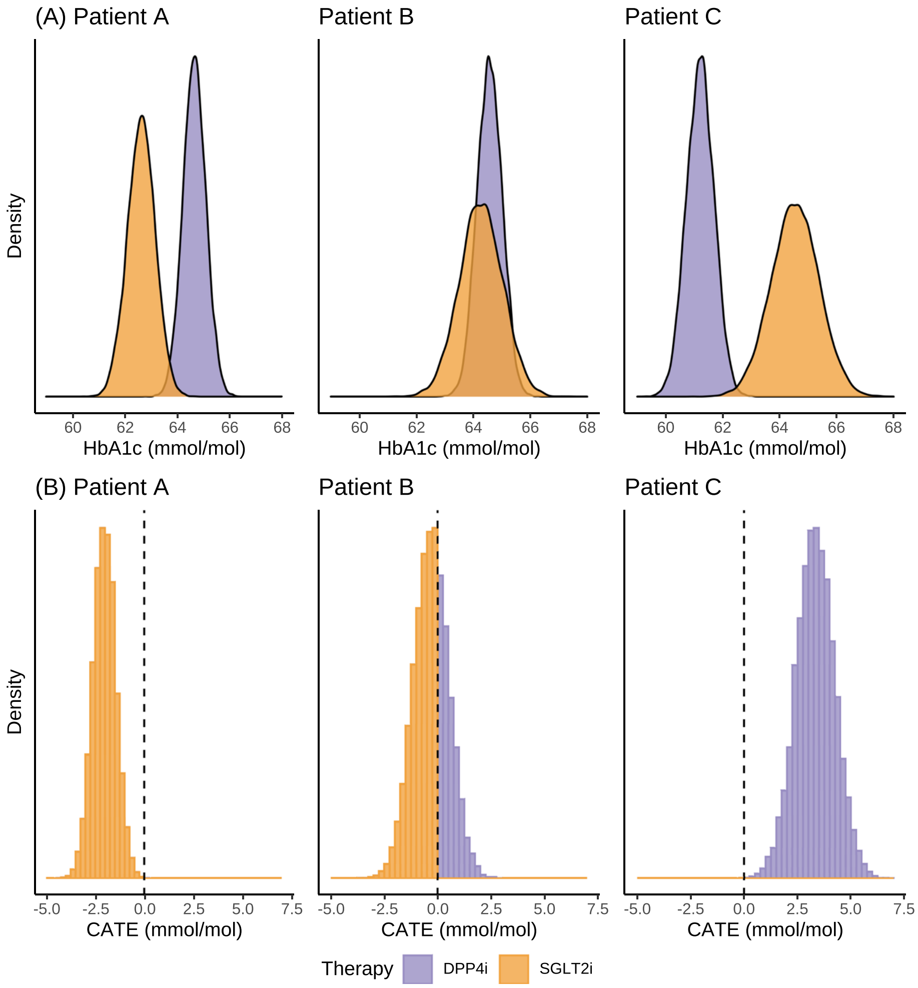

Fig. 2 **A** Posterior predictive distributions for HbA1c outcome at 6-month post-treatment for three synthetic but representative patients. For patient A, SGLT2i has a 98% probability of performing better than DPP4i. **B** Predicted treatment response difference or conditional average treatment effect (CATE) at 6-months. A negative value corresponds to a benefit on SGLT2i, and a positive value corresponds to a benefit on DPP4i. For patient B, SGLT2i has a 58% probability of performing better than DPP4i. For patient C, DPP4i has a >99% probability of performing better than SGLT2i. Patient [A,B,C]: Number of Past Drugs [4,3,4], Number of Current Drugs [2,2,2], HbA1c [67,75,65], eGFR [84.2,66.6,67.9], ALT(log) [3.4,2.8,2.6], BMI [26.1,33.4,28.5], Age [68,79,81]

of DPP4i. Patient B is predicted a possible average HbA1c benefit between 0–3 mmol/mol for both therapies. For patient C, there is a high probability of DPP4i resulting in a greater HbA1c reduction than SGLT2i, with a likely 1–7 mmol/mol benefit on average. As discussed earlier, these are posterior predictive distributions for the *mean* response for a patient with each set of characteristics (see Table 3).

Predictions for patients with missing data using the Bayesian treatment selection model

The inclusion of DPMM in the Bayesian model enables it to make predictions for patients with missing information. These predictions are possible by averaging over the conditional posterior predictive distributions for the missing values, given the existing data, and the uncertainty of the distributions is in part related to the degree of missingness. Figure 3 demonstrates how the data available for patient A influences the uncertainty of the posterior predictive distributions for the missing variables and the response. For each scenario, different sets of predictor variables are artificially missing. Each row displays how the uncertainty of the posterior predictive distributions for the missing covariates change as more data becomes available. The uncertainty of the posterior predictive treatment response distribution decreases as more variables become available. As baseline HbA1c is the most influential predictor in

Table 3 Posterior predictive distributions

In Bayesian modelling, it is possible to derive a full *posterior predictive distribution* for any new individual. Thus the uncertainty associated with a prediction is captured through a probability distribution, from which point estimates can be derived, or alternatively, probabilistic questions can be asked, such as those described in the Probabilistic predictions using the Bayesian treatment selection model section. The definition of a posterior predictive distribution for an individual with complete predictor information is:

$$f\left(Y_* \mid \mathbf{X}_*, \mathbf{X}, \mathbf{Y}\right) = \int_{\boldsymbol{\Psi}} f\left(Y_* \mid \mathbf{X}_*, \boldsymbol{\psi}\right) f\left(\boldsymbol{\psi} \mid \mathbf{X}, \mathbf{Y}\right) d\boldsymbol{\psi}. \tag{4}$$

Thus, the predictive distribution *integrates* (or *averages*) over the posterior distribution for the parameters and thus naturally incorporates the uncertainties about the parameters as well as those arising from the underlying model. The examples in this paper present the posterior predictive distribution for the expected outcome, with a given set of characteristics defined by:

$$f\left(E\left(Y_*\right) \mid \mathbf{X}_*, \mathbf{X}, \mathbf{Y}\right) = \int_{\boldsymbol{\Psi}} E\left(Y_* \mid \mathbf{X}_*, \boldsymbol{\psi}\right) f\left(\boldsymbol{\psi} \mid \mathbf{X}, \mathbf{Y}\right) d\boldsymbol{\psi}. \tag{5}$$

We are free to use either distribution (4–5), dependent on the context, although in practice, most treatment selection models use an analogous approach to (5), focusing on the conditional average treatment response. Please see the Supplementary Material for more information about how different predictive distributions can be derived and sampled from (including when predictors are missing, in which case we integrate the missing information modelled using the DPMM).

the model, the availability of the baseline HbA1c value vastly reduces the uncertainty of the prediction compared to other variables.

Variable influence on the response prediction using the Bayesian treatment selection model

Since the model provides posterior predictive distributions for the missing data, a new question can be proposed—which missing variable will, if collected, give the most information about the treatment response? As an example, consider a patient with only three predictors available (number of previous therapies, current therapies and age). With these three predictors, the model predicts similar treatment responses for both DPP4i and SGLT2i. For each missing predictor, a range of treatment response distributions can be estimated for different values of the predictors, which have been chosen as different quantiles from the corresponding conditional posterior predictive distributions for the missing variables given the observed variables. This provides insight into how the treatment response prediction might be affected across a range of plausible predictor values.

Figure 4 showcases the conditional probability distributions for any missing variable, and we can explore what we might expect to see at different values informed by these distributions. For example, the treatment response changes drastically depending on starting HbA1c. For this patient, if HbA1c was equal to the 5% quantile value, there is a high probability the expected treatment response would be below 60 mmol/mol with DPP4i therapy but a much lower probability of being below the same threshold under SGLT2i. On the other hand, if HbA1c was equal to the 50% quantile value, then the expected HbA1c response is between 60–70 mmol/mol for both treatments, with considerable overlap between the treatment response predictions for both therapies. Lastly, if HbA1c was equal to the 95% quantile value, there is a low probability of the expected treatment response being below 70 mmol/mol for DPP4i but a higher probability of being below the same threshold for SGLT2i. In contrast, when other missing variables take different quantile values, it results in very similar treatment responses across all quantiles. All of the above provides evidence that, given the choice, starting HbA1c would be the most useful additional variable to collect to improve the choice of therapy for this patient.

High-risk patient prediction using the Bayesian treatment selection model

In addition to all previous aspects, the Bayesian model could also be used in a different way. For example, in the case of a high-risk patient, it may be desirable to choose the treatment response that is most likely to result in a reduction of HbA1c to below 60 mmol/mol. Hence we could choose treatment on the basis of overall predicted benefit, or we could select the therapy with the highest probability of meeting some target threshold. Using Fig. 5 as a hypothetical example, there is a high probability of SGLT2i resulting in a better treatment response than DPP4i for this patient. In a conventional approach, SGLT2i would be the therapy chosen. However, it is critical to note that SGLT2i has a higher probability of

(See figure on next page.)

Fig. 3 Predictive distributions for missing data and treatment response conditional on the scenario for a patient at 6-months. Unobserved true values are given by the black vertical line and its prediction by the grey distribution. As more variables become known, the uncertainty in the predictions decreases. Scenario 1: Past Drugs, Current Drugs, Age known. Scenario 2: Past Drugs, Current Drugs, Age, HbA1c known. Scenario 3: Past Drugs, Current Drugs, Age, HbA1c, BMI known. Scenario 4: Past Drugs, Current Drugs, Age, HbA1c, BMI, eGFR known. Scenario 5: All variables known. Patient: Number of Past Drugs [2], Number of Current Drugs [0], HbA1c [73], eGFR [68.9], ALT(log) [2.9], BMI [30.1], Age [75]

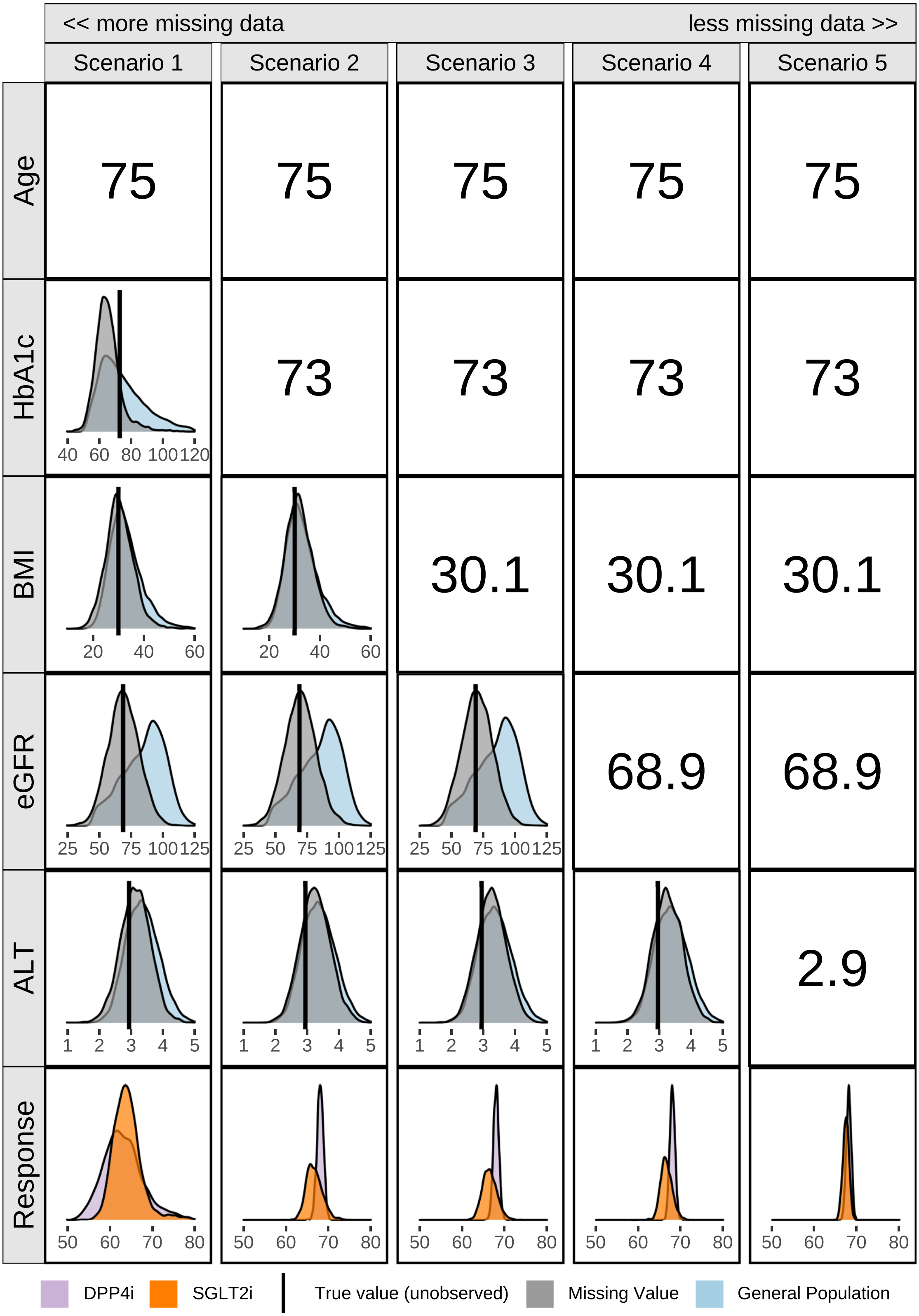

Fig. 3 (See legend on previous page.)

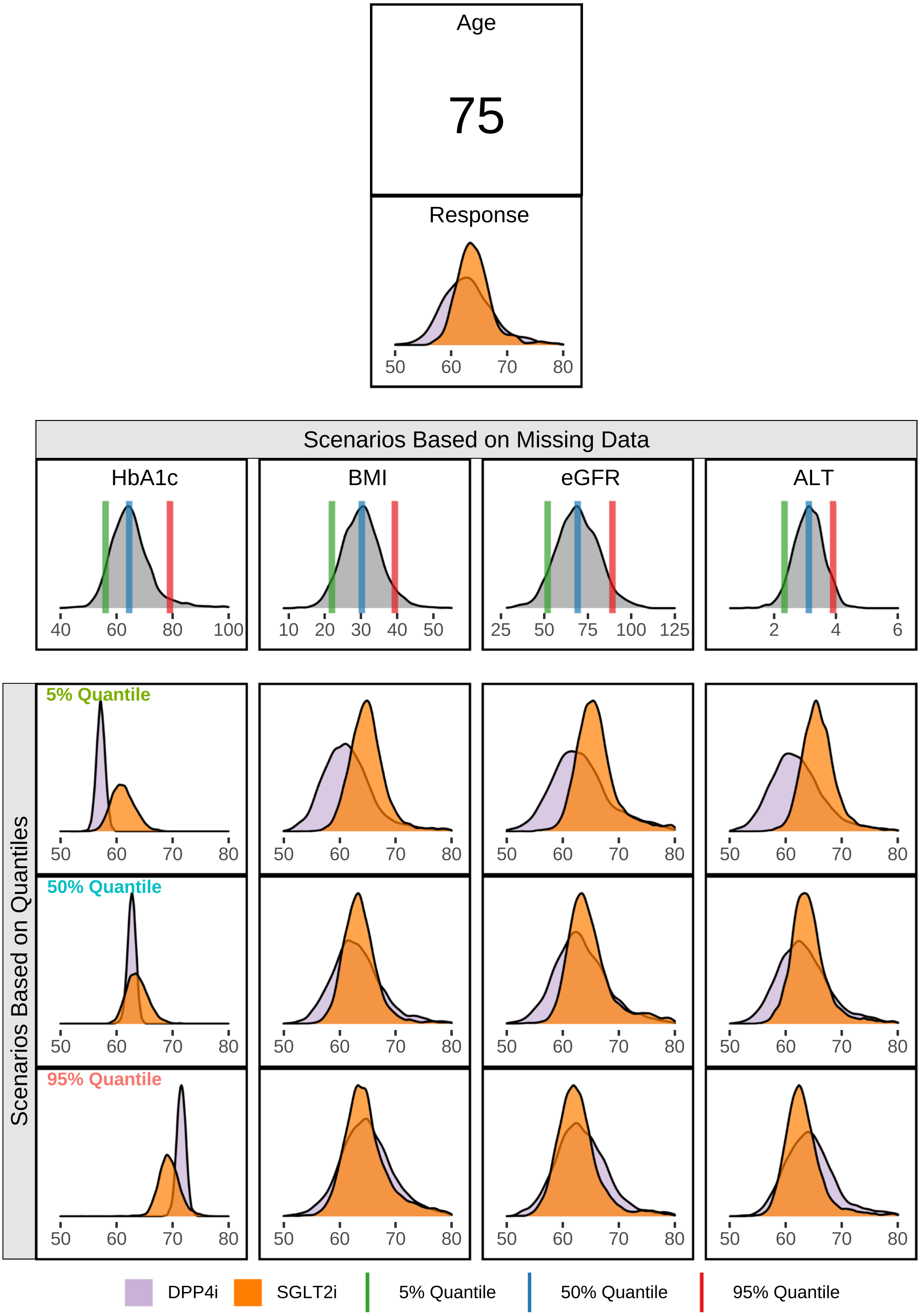

Fig. 4 Analysis of variable influence on response prediction for a patient at 6-months. The initial response prediction is conditional on age, number of previous and current therapies. Further predictions are conditional on quantile values for HbA1c, BMI, eGFR and ALT. Out of the 4 missing covariates, the collection of HbA1c will, in theory, provide the most information about treatment response for the patient. Patient: Number of Past Drugs [2], Number of Current Drugs [0], Age [75]

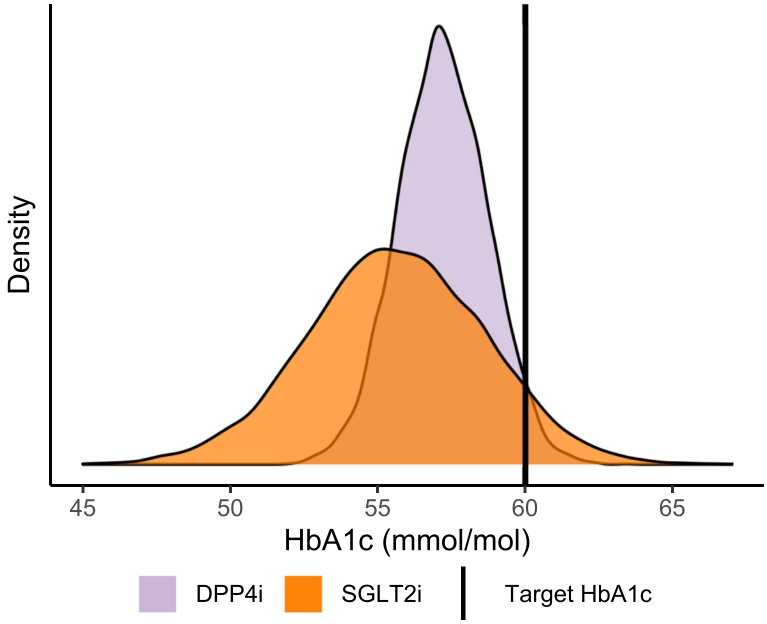

Fig. 5 Therapy probability distribution at 6-month prediction for an arbitrary patient. The patient has a <60 mmol/mol target HbA1c outcome. In this situation, SGLT2i has a 7.7% chance of >60 mmol/mol treatment response, whereas DPP4i has a 3.1% chance of >60 mmol/mol. Due to the lower probability of treatment response above the target, DPP4i is the right therapy for this patient

resulting in a value of HbA1c above 60 mmol/mol than DPP4i. In this instance, the fact that DPP4i is less likely to result in a value above the target could be more influential in the choice of therapy than SGLT2i being more likely to result in lower treatment response on average. Choosing DPP4i as a therapy for this patient gives the highest probability of achieving an average treatment response below the target.

Discussion

This paper provides details on how to fit a Bayesian treatment selection model with continuous and categorical covariates using a DPMM to capture the joint-covariate distribution. The model enables model fitting and response prediction for patients with complete and incomplete data (under the assumption that missingness is MCAR/MAR), similarly to [14, 27, 39]. It can be thought of as an alternative to Bayesian profile regression [11, 12] in which the outcome model-of-interest is conditioned explicitly on the covariates, rather than on the cluster memberships derived from the DPMM. This makes it more directly applicable to augment existing models. Crucially, it can be used to gain insight into the utility of collecting further data on patients with incomplete predictors in clinical practice.

The Bayesian treatment selection model presented in this paper augments the classical counterfactual model developed by Dennis et al. [2]. The model validation is carried out in an analogous way to the original model. This validation assesses whether the treatment selection model accurately predicts the optimal treatment rather than the exact therapy response. With that in mind, the performance of the Bayesian model is comparable to [2], even after augmenting the data to include the incomplete-case individuals (Supplementary Table S1 and Fig. S5). The regression parameter values are similar between the three models, with the Bayesian model achieving similar regularisation through the use of weakly informative regularising priors [40]. Including individuals with incomplete data in the model fitting process reduces the uncertainty of regression parameters marginally (Supplementary Fig. S5). This is because here the original data set was very large, and the number of individuals with incomplete data was a relatively small proportion of the data set (especially since most individuals would only be missing a few covariates at most). Hence, in this case, the missing data do not influence the posterior distributions for the parameters very much. However, this may not be true in other settings, and this approach allows a direct comparison of the difference between models fitted to complete-case data and incomplete-case data. More importantly, the treatment selection model now has a mechanism to produce predictions for new individuals with incomplete information, thus facilitating its use in clinical practice.

There are several additional benefits of this type of approach. In this case, we can estimate the treatment response alongside the difference in treatment response for both therapies. Furthermore, due to the inclusion of the DPMM, the model can provide predictive distributions for any missing covariates [14, 27, 39]. The availability of predictive distributions for missing covariates, in turn, can be used to influence the choice of future covariate data collection to improve treatment response prediction. Instead of having a set rule, we could use these distributions (Fig. 4) to choose to collect a variable that reduces the uncertainty in the predictive distribution for the difference between the two treatments. Alternatively, we could choose to collect a variable that optimises the posterior probability of preference for one therapy over the other. Moreover, since we obtain joint predictive distributions for all the missing variables, these can be utilised to look at the changes in expected treatment response for multiple variables simultaneously (although in Fig. 4 we show the univariate *marginal* predictive distributions for each variable, averaging over the others).

Furthermore, even in the presence of complete data, the inclusion of the DPMM in the model building enables the potential to make predictions for patients with incomplete predictor information, which is useful in the deployment of prediction models in clinical practice, where it may not be possible to collect all predictor information for a new patient. This is a novel approach with considerable advantages for prediction from treatment selection models, and many of these advantages would also be true for standard clinical prediction models. We have implemented this model in standard software (NIMBLE), and as such the DPMM can be used to augment any regression model supported by NIMBLE (for example, the same ideas could be applied to e.g. survival models, logistic regression and so on), without requiring new inference algorithms to be developed.

Currently, the assumption that missing predictor data are MCAR/MAR means that it is not necessary to explicitly model the missingness mechanism (as long as the parameters for the missingness model are *a priori* independent of the other parameters—in which case the missingness mechanism can be analytically integrated out, leaving the posterior defined in (2)). However, there are some restrictions. DPMMs, like many non-parametric approaches, are not good models for extrapolation. This means that for the DPMM to be a good representation of the joint distribution of the predictors, there needs to be some observed data throughout all regions of non-negligible density. As an extreme example, assume there are multiple predictors, $X_1, X_2, \ldots, X_p$, where X_1 is *truncated*, based on X_2 (so X_1 is missing for any $X_2 > x^*$ say). Here the missingness mechanism for X_1 depends only on X_2, so the data are MAR. However, since there are no samples for X_1 in the region $X_2 > x^*$ the DPMM will be unable to adequately estimate the conditional distribution $f(X_1 \mid X_2 > x^*)$ on its own (although in hierarchical prediction models the response variable Y can also play a role in constraining the distributions for the missing data). In this case, we may need a more structured model for $X_1 \mid X_2$ if we wish to extrapolate into regions with no available data; although note that we can consider other hierarchical structures, such as $f(X_1 \mid X_2) f\left(X_2, \ldots, X_p\right)$ where a DPMM is used to model the density $f\left(X_2, \ldots, X_p\right)$. In this setting, despite major challenges (large sample size, large absolute numbers of missing information, reasonable numbers of covariates, clear non-normality in the data, combinations of continuous and categorical predictors) we were able to fit the described models in ≈24 hours. We note that this is substantially more than the complete-case model (3 hours), but it would be expected that the approach would be computationally challenging to fit for large datasets with lots of missing data and/or a high number of predictors. Despite this, once the models are fitted then predictions will remain relatively quick to obtain. On the other hand, a complete case analysis is straightforward to perform, only requiring the NIMBLE code we provide without the need for the custom MCMC samplers.

The further development of DPMM models is an active area of research, and some extensions that could be of interest in future work include exploring whether a more complex model structure, such as those previously discussed based on the joint modelling of the response and predictor variables [27] could improve predictive performance. The predictive performance could also be compared to simpler imputation techniques such as multivariate imputation by chained equations (MICE) (see [41] for more information). A key constraint in making these models straightforward and cost-effective to employ in clinical practice is ensuring that they are parsimonious, and hence this is an area where often variable selection methodology is employed, and although software exists to do this within DPMMs, such as the PReMiuM package in R [12], this currently only works for profile regression, or if a DPMM is used as a joint model for the response and predictors. There is published work exploring ways to handle missing data in prediction models when the missingness mechanism is MNAR (see e.g. [39, 42]), which may be relevant for other future applications. One potential advantage of separating the response model from the predictor distribution model is that cutting-edge treatment selection approaches such as Bayesian Causal Forests [43] could be used for the former, alongside a DPMM for the latter, but this would require more methodological development.

As a final note, we found that the DPMM structure employed here did a good job of capturing the covariate distribution without overfitting (Fig. 1 and Supplementary Fig. S4), but it is worth considering that there are many alternative ways to specify joint distributions with dependency structures amongst the predictors which may be applicable in other situations (see [44] for a review).

Conclusion

In this work, we develop a hierarchical joint model using a spline-based treatment selection model for type 2 diabetes, alongside a DPMM as a flexible way of modelling the complex multivariate distribution of predictor variables. This facilitates both model inference and prediction in the presence of incomplete information. The inclusion of the DPMM in the joint model unlocks information that can be used by practitioners during the decision-making process. This modelling technique is being applied in treatment selection models for type 2 diabetes but could provide benefits to other types of prediction models.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-023-02400-3.

Additional file 1.

Acknowledgements

This article is based on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare products Regulatory Agency. CPRD data is provided by patients and collected by the NHS as part of their care and support.

MASTERMIND consortium

Prof Andrew Hattersley[1], Prof Ewan Pearson[2], Dr Angus Jones[1], Dr Beverley Shields[1], Dr John Dennis[1], Dr Lauren Rodgers[1], Prof William Henley[1], Prof Timothy McDonald[1], Prof Michael Weedon[1], Prof Nicky Britten[1], Catherine Angwin[1], Dr Naveed Sattar[3], Dr Robert Lindsay[3], Prof Christopher Jennison[4], Prof Mark Walker[5], Prof Kennedy Cruickshank[6], Dr Salim Janmohamed[7], Prof Christopher Hyde[1], Prof Rury Holman[8], Prof Andrew Farmer[8], Prof Alastair Gray[8], Prof Stephen Gough[8], Dr Olorunsola Agbaje[8], Dr Trevelyan McKinley[1], Dr Sebastian Vollmer[9], Dr Bilal Mateen[7], Prof William Hamilton[1]

[1]University of Exeter, Exeter, England
[2]University of Dundee, Dundee, Scotland
[3]University of Glasgow, Glasgow, Scotland
[4]University of Bath, Bath, England
[5]University of Newcastle, Newcastle upon Tyne, England
[6]Kings College London, London, England
[7]University College London, London, England
[8]University of Oxford, Oxford, England
[9]University of Kaiserslautern, Kaiserslautern, Germany.

Authors' contributions

PC, JMD, JB, BS and TJM conceived and designed the study. PC, with support from JMD and TJM analysed the data and developed the code. All authors contributed to the writing of the article, provided support for the analysis and interpretation of results, critically revised the article, and approved the final article.

Funding

This research was funded by the Medical Research Council (UK) (MR/N00633X/1). P.C., T.J.M. and J.B. are funded by Research England's Expanding Excellence in England (E3) fund. J.M.D. is supported by an Independent Fellowship funded by Research England's Expanding Excellence in England (E3) fund.

Availability of data and materials

The routine clinical data analysed during the current study are available in the CPRD repository (CPRD; https://cprd.com/research-applications), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. For re-using these data, an application must be made directly to CPRD. A synthetic sample data is available on GitHub within the repository "PM-Cardoso/DPMM-tsm".

Declarations

Ethics approval and consent to participate

Approval for CPRD data access and the study protocol was granted by the CPRD Independent Scientific Advisory Committee (Project–ID: ISAC 13_177R), and informed consent was waived for this retrospective study. Individual patients can opt out of sharing their data for CPRD, and CPRD does not collect data for these patients. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]University of Exeter, Medical School, Exeter, England.

Received: 25 January 2023 Accepted: 11 December 2023
Published online: 08 January 2024

References

1. Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med. 2020;172(35).
2. Dennis JM, Young KG, McGovern AP, Mateen BA, Vollmer SJ, Simpson MD, et al. Development of a treatment selection algorithm for SGLT2 and DPP-4 inhibitor therapies in people with type 2 diabetes: a retrospective cohort study. Lancet Digit Health. 2022;4(12):e873–83.
3. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley; 2002.
4. McLachlan GJ, Rathnayake S, Lee SX. Comprehensive Chemometrics: Chemical and Biochemical Data Analysis. 2nd ed. Oxford: Elsevier; 2020.
5. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
6. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40–9.
7. Sisk R, Sperrin M, Peek N, van Smeden M, Martin GP. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. Stat Methods Med Res. 2023;32(8):1461–77.
8. Moons KGM, Donders RART, Stijen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092–101.
9. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. New York: Chapman & Hall/CRC; 2013.
10. McAuliffe JD, Blei DM, Jordan MI. Nonparametric empirical Bayes for the Dirichlet process mixture model. Stat Comput. 2006;16:5–14.
11. Molitor J, Papathomas M, Jerrett M, Richardson S. Bayesian profile regression with an application to the national survey of children's health. Biostatistics. 2010;11(3):484–98.
12. Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S. PReMiuM: An R package for profile regression mixture models using Dirichlet processes. J Stat Softw. 2015;64(7):1–30.
13. Banerjee A, Murray J, Dunson D. Bayesian learning of joint distributions of objects. Artif Intell Stat. 2013;31:1–9.

14. DeYoreo M, Reiter JP, Hillygus DS. Bayesian mixture models with focused clustering for mixed ordinal and nominal data. Bayesian Anal. 2017;12(3):679–703.
15. Akande O, Li F, Reiter J. An empirical comparison of multiple imputation methods for categorical data. Am Stat. 2017;71(2):162–70.
16. Ferguson TS. A Bayesian analysis of some nonparametric problems. Annals Stat. 1973;1(2):209–30.
17. Favaro S, Walker SG. A generalized constructive definition for the Dirichlet process. Stat Probab Lett. 2010;78(16).
18. Peel D, McLachlan G. Finite Mixture Models. New York: Wiley; 2000.
19. Papaspiliopoulos O, Roberts GO. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika. 2008;95(1):169–86.
20. Daniels MJ, Linero AR, Roy J. Bayesian Nonparametrics for Causal Inference and Missing Data. Chapman & Hall/CRC; 2023.
21. Li Y, Schofield E, Günen M. A tutorial on Dirichlet process mixture modeling. J Math Psychol. 2019;91:128–44.
22. Zio MD, Guarnera U, Luzi O. Imputation through finite Gaussian mixture models. Comput Stat Data Anal. 2007;51(11):5305–16.
23. Kim HJ, Reiter JP, Wang Q, Cox LH, Karr AF. Multiple imputation of missing or faulty values under linear constraints. J Bus Econ Stat. 2014;31(2):375–86.
24. Si Y, Reiter JP. Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. J Educ Behav Stat. 2013;38(5):499–521.
25. Wang C, Liao X, Carin L, Dunson DB. Classification with incomplete data using Dirichlet process priors. J Mach Learn Res. 2010;11(12).
26. Manrique-Vallier D, Reiter JP. Bayesian simultaneous edit and imputation for multivariate categorical data. J Am Stat Assoc. 2017;112(520):1708–19.
27. Roy J, Lum KJ, Zeldow B, Dworkin JD, Lo Re III V, Daniels MJ. Bayesian nonparametric generative models for causal inference with missing at random covariates. Biometrics. 2018;74(4):1193–202.
28. Wade S, Mongelluzzo S, Petrone S. An enriched conjugate prior for Bayesian nonparametric inference. Bayesian Anal. 2011;6(3):359–86.
29. Wade S, Dunson DB, Petrone S, Trippa L. Improving prediction from Dirichlet process mixtures via enrichment. J Mach Learn Res. 2014;15:1041–71.
30. Müller P, Erkanli A, West M. Bayesian curve fitting using multivariate normal mixtures. Biometrika. 1996;83(1):67–79.
31. Dennis J. Precision medicine in type 2 diabetes: using individualized prediction models to optimise selection of treatment. Diabetes. 2020;69:2075–85.
32. de Valpine P, Turek D, Paciorek CJ, Anderson-Bergman C, Temple Lang D, Bodik R. Programming with models: writing statistical algorithms for general model structures with NIMBLE. J Comput Graph Stat. 2017;26:403–13.
33. de Valpine P, Paciorek C, Turek D, Michaud N, Anderson-Bergman C, Obermeyer F, et al. NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling. 2022. R package version 0.12.2. https://cran.r-project.org/package=nimble.
34. R Core Team. R: a language and environment for statistical computing. Vienna, Austria. 2021. https://www.R-project.org/.
35. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data resource profile: clinical practice research datalink (CPRD). Int J Epidemiol. 2015;44(3):827–36.
36. Harrell Jr FE. Regression Modeling Strategies. New York: Springer International Publishing; 2015.
37. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci. 1992;7(4):457–511.
38. Bossuyt PM, Parvin T. Evaluating biomarkers for guiding treatment decisions. EJIFCC. 2015;26(1):63–70.
39. Linero AR, Daniels MJ. A flexible Bayesian approach to monotone missing data in longitudinal studies with nonignorable missingness with application to an acute schizophrenia clinical trial. J Am Stat Assoc. 2015;110(509):45–55.
40. Bürkner PC. brms: an R package for Bayesian multilevel models using Stan. J Stat Softw. 2017;80(1):1–28.
41. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1–67.
42. van Hasselt M. Bayesian inference in a sample selection model. J Econ. 2011;165(2):221–32.
43. Hahn PR, Murray JS, Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). Bayesian Anal. 2020;15(3):965–1056.
44. Daniels MJ, Gaskins JT. Bayesian methods for the analysis of mixed categorical and continuous (incomplete) data. In: de Leon AR, Chough KC, editors. Analysis of Mixed Data: Methods and Applications. Chapman & Hall/CRC; 2013.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.