PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Review

Cite this article: Müller P, Chandra NK, Sarkar A. 2023 Bayesian approaches to include real-world data in clinical studies. *Phil. Trans. R. Soc. A* **381**: 20220158.
https://doi.org/10.1098/rsta.2022.0158

Received: 6 August 2022
Accepted: 27 September 2022

One contribution of 16 to a theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

Subject Areas:
statistics

Keywords:
common atoms mixture model, propensity scores, real-world data

Author for correspondence:
P. Müller
e-mail: pmueller@math.utexas.edu

Bayesian approaches to include real-world data in clinical studies

P. Müller[1], N. K. Chandra[2] and A. Sarkar[1]

[1]Department of Statistics and Data Sciences, The University of Texas at Austin, 2317 Speedway D9800, Austin, TX 78712-1823, USA
[2]Department of Mathematical Sciences, The University of Texas at Dallas, 800 W Campbell Road, Richardson, TX 75080-3021, USA

PM, 0000-0002-2948-1229

Randomized clinical trials have been the mainstay of clinical research, but are prohibitively expensive and subject to increasingly difficult patient recruitment. Recently, there is a movement to use real-world data (RWD) from electronic health records, patient registries, claims data and other sources in lieu of or supplementing controlled clinical trials. This process of combining information from diverse sources calls for inference under a Bayesian paradigm. We review some of the currently used methods and a novel non-parametric Bayesian (BNP) method. Carrying out the desired adjustment for differences in patient populations is naturally done with BNP priors that facilitate understanding of and adjustment for population heterogeneities across different data sources. We discuss the particular problem of using RWD to create a synthetic control arm to supplement single-arm treatment only studies. At the core of the proposed approach is the model-based adjustment to achieve equivalent patient populations in the current study and the (adjusted) RWD. This is implemented using common atoms mixture models. The structure of such models greatly simplifies inference. The adjustment for differences in the populations can be reduced to ratios of weights in such mixtures.

This article is part of the theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

1. Introduction

Randomized controlled trials (RCT) remain the gold standard for clinical studies to estimate the effect of an intervention. However, RCTs, while critical to get drugs

THE ROYAL SOCIETY PUBLISHING

to the market are beset with rising costs and difficulties in volunteer recruitment. Prolonged duration of RCTs has impeded introduction of much needed therapies to consumers. Against the background of these challenges the US *21st Century Clinical Care* act was passed in 2016, to encourage the use of readily available real world data (RWD) in drug development. This legislation and the increased availability of RWD have paved the way for RWD in clinical trial designs. Here RWD is defined as data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. The resulting real-world evidence (RWE) is the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of such RWD. Pursuant to the *21st Century Clinical Care* act, the U.S. Food and Drug Administration (FDA) is working with stakeholders to understand how RWE can best be used to increase the efficiency of clinical research, as outlined, for example in [1].

One of the main issues in using RWD to complement RCTs is the need to demonstrate equivalence of the patient populations under comparison, or achieve such equivalence by appropriate adjustment or pre-processing. Propensity scoring is perhaps the most widely used approach for matching populations to achieve equivalence. Propensity scores (PS) are often estimated by logistic regression. However, in high dimensions, the inclusion of interaction terms incurs an explosion of the design matrix. For this and related reasons, high dimensionality of covariates, mixed data types and data missingness are fatal limitations to compare populations for equivalence. It turns out that a preponderance of covariates are categorical. So, dimension reduction by standard principal components (PCA) is not immediately feasible and alternate methods are necessary. As an alternative to PS-based methods, in this article we also review a model-based approach using common atoms mixture models. One motivation for this choice is the observation that a variety of disease mechanisms and epidemiological factors induce population heterogeneity, making it natural to use mixture models to account for intrinsic heterogeneity and to estimate component-specific parameters for model-based inference.

In this article, we review several recently proposed methods using model-based Bayesian inference to incorporate RWD in clinical trial design, including the mentioned PS-based methods. We discuss in more detail a recently developed method introduced in [2] which uses common atoms mixture models to account for heterogeneity. The model is parametrized so as to allow parsimonious evaluation of weights for adjustment to achieve equivalent populations, or—equivalently—to allow easy and flexible estimation of PS. These methods break the distinction of model-based versus PS-based methods, as they can be seen as both. The scheme implements stochastic stratification by PS, that is, inference is averaged with respect to uncertainties in the stratification. This happens naturally in the proposed mixture model framework. We develop the model and inference for the following context. Assume an investigator is designing a new study (current study) and wishes to make use of RWD to supplement the current single-arm treatment-only study with a synthetically created control arm. Here, the RWD could be data from earlier studies (historical studies), electronic health records or repositories. We assume a typical phase II study, with event time or binary endpoint, as it arises, for example, in phase II trials for glioblastoma (GBM) patients without controls. See, for example, [3] for a discussion of the reasons that led investigators to design single-arm trials in GBM.

While the discussion is in the context of this scenario, the approach is more widely applicable. It can be used to create equivalent populations in any other context. For example, it could be used to compare two treatments based on RWD, assembling both treatment arms from the same RWD source. Or it could be used for interim analysis and stopping decisions [4].

Inference is implemented using Markov chain Monte Carlo (MCMC) posterior simulation, which since the seminal work of A.F.M. Smith and collaborators in the 1990s [5,6] has enabled researchers to use increasingly more complex inference models. At the core of the approach is a common atoms mixture (CAM) model for the two patient populations, including patients in the current single-arm trial and patients in the RWD. An interesting aspect of posterior inference in the proposed CAM is that appropriately reweighed posterior simulation under one mixture submodel can be used to evaluate expectations with respect to the other mixture submodel. The CAM model is an example of a non-parametric Bayesian (BNP) model, whose wide support, i.e.

the ability to fit virtually any distribution, is critical for use in this application. Some of the early work of A.F.M Smith was instrumental in introducing BNP models in Bayesian inference [7,8].

In §2, we review some currently used methods. The CAM model and its use for the construction of synthetic control arms are introduced in §3. In §4, we show an application to single arm, treatment-only phase II studies for glioblastoma patients, and results from a simulation study to compare the proposed method with alternatives based on propensity scores.

2. Real-world evidence in Bayesian clinical trial designs

Several methods have been proposed in the literature to incorporate information from RWD in Bayesian clinical trial design. Several early methods consider the particular case of using information from historical studies to formulate informative priors. These include historical data priors, commensurate priors and meta-analytic priors, as briefly reviewed below. Several other methods are based on PS for patients with given baseline covariates. The PS can be for treatment selection, or for a patient being selected into one or the other study. Some methods are based on flexible non-parametric regression models for potential outcomes (under treatment and control). Another commonly used element of different methods is the use of patient-specific weights to either thin out datasets or implement analyses with such weights. Details of these methods are discussed in the following subsections.

The following review includes methods developed for a variety of different setups involving RWD at different levels. In particular, some methods are meant for use in studies with a concurrent control when RWD is used to augment this (often reduced sample size) control arm. Other methods assume that a study is restricted to inference under a single condition. Yet other methods create a synthetically constructed control arm to supplement a single-arm, treatment-only study. In particular, methods discussed under (a) and (d) below have no notion of constructing synthetic cohorts, but are suitable when either a concurrent control is available, or inference is about a single condition only. The approach in (b) is explicitly about augmenting a concurrent control arm, while (c) allows for both use cases. Relatedly, the broad definition of RWD (compare the introduction) is reflected in a wide variety of use cases, including applications where RWD refers to historical studies, registry data, observational studies or insurance data.

(a) Historical data priors

Under the Bayesian paradigm [9], a natural way of exploiting external information is by way of informative prior probability models. If the external information are historical trials, one widely used method to develop such priors is the use of the so-called power priors. The construction and use of power priors is reviewed in [10]. We define a prior probability density (or function) to be proportional to a fractional power of the likelihood function for the historical data. The fractional power formalizes a notion of discounting the information from the earlier study. Let D denote the data for the current study, and D_1 the data from a historical study. Let $p(D_1 \mid \theta)$ denote the likelihood function for the historical data, assuming a model indexed by an unknown parameter θ, and let γ_1 denote the fractional power. Assume then that the inference model $p(D \mid \theta)$ for the current data is indexed by the same parameter θ. That is, the inference model for the historical and the current study share the same parameter θ. The historical data prior is then defined as

$$\pi(\theta \mid D_1, \gamma_1) = \frac{1}{c(\gamma_1)}\, \pi(\theta) \cdot p(D_1 \mid \theta)^{\gamma_1},$$

with normalization constant $c(\gamma_1)$. Using fixed γ_1, this is known as the conditional power prior. Extending the model to allow for uncertainty on γ_1 defines the *modified joint power prior*

$$\pi(\theta, \gamma_1 \mid D_1) \propto \frac{\pi(\theta) \cdot p(D_1 \mid \theta)^{\gamma_1}}{c(\gamma_1)}\, \pi(\gamma_1). \tag{2.1}$$

The implied complete conditional distribution $p(\gamma_1 \mid \theta, D)$ reveals one limitation of the modified power prior. The conditional $p(\gamma_1 \mid \theta, D)$ is free of the current data D. This has led to the development of the *commensurate power prior* by Hobbs *et al.* [11]. The commensurate prior allows for the possibility of different parameters for the historical and current study. Let θ_1 denote the parameter for the inference model $p(D_1 \mid \theta_1)$ for the historical data, and let θ denote the parameter in $p(D \mid \theta)$, as before. The commensurate prior links the two inference models by postulating a hierarchical prior probability model with positive prior probability for $\theta \approx \theta_1$. Let $N(\cdot \mid m, V)$ denote a normal p.d.f. with moments (m, V). A commensurate prior is defined as

$$\pi(\theta, \gamma_1, \tau) \propto \left\{ \int \frac{p(D_1 \mid \theta_1)^{\gamma_1}}{c(\gamma_1)} \cdot N\left(\theta \mid \theta_1, \frac{1}{\tau}\right) d\theta_1 \right\} \pi(\gamma_1 \mid \tau)\, \pi(\tau).$$

The two parameters θ, θ_1 are linked by assuming $\theta \mid \theta_1, \tau \sim N(\theta_1, 1/\tau)$. The normal precision τ is known as the commensurability parameter. The model can adapt to prior-data conflict if needed. Of course, alternatives to the normal model for $p(\theta \mid \theta_1, \tau)$ could be used if desired.

Considering a setting with multiple prior studies [12] define the meta-analytic prior (MAP). Let $s = 1, \ldots, S$ index multiple prior (external) studies. Similar to before, we assume inference models indexed by $(\theta_s, s = 1, \ldots, S)$ for the historical data and θ for the current data, respectively. The MAP prior is defined as the posterior predictive distribution for θ in a hierarchical extension that includes submodels for each of the historical studies and the current study, as

$$\pi(\theta \mid D_s, s = 1, \ldots, S) \propto \int p(\theta \mid \eta) \left\{ \prod_s \int p(D_s \mid \theta_s)\, dp(\theta_s \mid \eta) \right\} \pi(\eta)\, d\eta.$$

In a default implementation [12] use a half-normal prior $\pi(\sigma^2)$ for across-study variance σ^2 in $p(\theta \mid \eta) = N(\theta \mid \eta_1, \eta_2 = \sigma^2)$. Similar to the commensurate prior, [13] define the *robust MAP* prior by introducing an additional component in the mixture model, allowing for θ and θ_s to be independent. Let $\widehat{\pi}(\theta)$ denote a mixture model approximation of the MAP prior, and let π_v denote a conjugate prior for the assumed sampling model $p(D \mid \theta)$. The robust MAP is defined as

$$\pi(\theta \mid D_1, \ldots, D_S) = (1 - w)\widehat{\pi}(\theta) + w\pi_v(\theta).$$

The mixture approximation $\widehat{\pi}$ in lieu of the actual MAP prior is introduced to simplify the model construction by defining the robust MAP as a simple extension of the same mixture. The marginal posterior on w characterizes the level of borrowing of strength.

(b) Non-parametric regression and potential outcomes

Using a potential outcome [14] framework, [15] develop a method to enrich an RCT with additional data on control using RWD. That is, they use RWD to supplement (a typically reduced sample size) control arm in a current study. They proceed by fitting a non-parametric regression model of outcomes as a function of baseline covariates for patients under control, and similarly for patients under treatment. The main innovation is that the earlier regression for outcomes under control includes a study indicator as an additional covariate, thus allowing probabilistic adjustment for study effects. The regression is implemented as a Bayesian random forest using Bayesian additive regression trees (BART) [16].

Formally, let $Y_i(a)$ denote the (potential) outcome for patient i under treatment a, let $S_i \in \{0, 1, \ldots, S\}$ denote an indicator for inclusion in the current study $(s = 0)$ or external data $(s > 0)$. Two random forest models for patients under treatment $A_i \in \{0, 1\}$ and covariates X_i define $(Y_i \mid A_i = 0, X_i = x, S_i = s) = f_0(x, s) + \epsilon_{0i}$ using a random forest $f_0(\cdot)$ and $(Y_i \mid A_i = 1, X_i = x, S_i = 0) = f_1(x) + \epsilon_{1i}$ with a second instance $f_1(\cdot)$ of the random forest model, and assumed i.i.d. residuals ϵ_{ai}. The model implies a well-defined conditional average treatment effect (CATE) as

$$\Delta = \frac{1}{N} \sum_i E\{Y_i(1) - Y_i(0) \mid \text{Data}, S_i = 0\}.$$

Here the BART models $f_a(\cdot)$ adjust for patient-level covariates, naturally allow for interactions, and capture potentially heterogeneous treatment effects across different data sources s.

(c) Propensity scores

Several methods use *propensity scores* (PS)

$$e(X) = p(S = 0 \mid X),$$

for current ($S = 0$) versus historical ($S = 1$) data, that is, $S \in \{0, 1\}$ is an indicator for a patient in a merged population being selected into one versus the other dataset. Wang *et al.* [17] stratify the data by equal quantiles of $e(X)$ in the current study. Let $r = 1, \ldots, R$ index these R quantiles. First, external data beyond this range is trimmed. Then let $f_{r0}(e)$ and $f_{r1}(e)$ denote density estimates for e restricted to stratum r in the current study and external data, respectively. Next, define a coefficient of similarity based on some distance measure (or measure of discrepancy) $\delta_r = \text{diff}\{f_{r0}, f_{r1}\}$. For inference under stratum r we then use a power prior with exponent $\gamma_r \propto \delta_r$, to report a stratum-specific treatment effect θ_r and an overall treatment effect $\theta = (1/R)\sum_r \theta_r$ (assuming equal quantiles).

Using a non-Bayesian approach, a similar scheme is developed in [18] defining a composite likelihood including factors for strata $r = 1, \ldots, R$. Let θ denote parameters of the outcome model. Let W_r^1 denote the indices of all patients in the RWD in statum r, and similarly W_r^0 for patients from the current study. A composite likelihood function is then defined as

$$L(\theta) = \prod_r \left\{ \prod_{i \in W_r^1} p(y_{1i} \mid \theta)^{\gamma_r} \prod_{i \in W_r^0} p(y_{0i} \mid \theta) \right\},$$

with $\gamma_r = \text{diff}\{f_{r0}, f_{r1}\}$, similar to the exponent γ_r in the stratified power prior.

The propensity-score integrated power prior [17] and the composite likelihood method [18] are implemented in the R package psrwe [19]. The program is easy to use, including functions to implement each step of the process, starting with one function to estimate the propensity scores $e(X)$, a function to create the stratification, and finally two functions to implement either the PS-integrated power prior or the composite likelihood, as desired. For the estimation of the propensity scores, the user can choose either logistic regression or random forests.

Many other uses of PS are developed in the recent literature. For example, [20] combine the use of PS and MAP priors. The approach evaluates PS $e(X)$ for selection into one or the other study, stratifies with respect to these propensity scores and then uses MAP priors to evaluate treatment effects in each stratum.

A generalization with multivariate propensity scores for selection of patients into one of $S > 2$ studies is developed in [21]. The method sets up a non-parametric Bayesian regression of outcomes on these multivariate propensity scores, and then evaluates an average treatment effect as a difference of fitted outcomes under treatment versus control.

(d) Patient-level weights

Golchi [22] comments on the nature of historical data priors as using study-specific weights, which appear as the fractional power in the historical data likelihood function. The use of such omnibus weights ignores the typically high level of heterogeneity of patient populations. The latter would more naturally justify patient-specific weights, with higher weights for patients who are more representative of the current patient population. Let then D_{si} denote data for the ith patient in study s with $s = 0$ for the current study and $s \geq 1$ for historical studies, and let $D_s = (D_{si}, i = 1, \ldots, n_s)$. Defining $\gamma_{si} = \delta_{\text{sim}}(D_{si}, D_0)$ as a similarity measure of D_{si} and D_0, we then proceed similar to historical data priors, but now with patient-specific weights to construct a prior $\pi(\theta) \propto \pi_0(\theta) \cdot \prod_{s \geq 1} p(D_{si} \mid \theta)^{\gamma_{si}}$. For parsimony, in a practical implementation replace γ_{si} by $\widetilde{\gamma}_{si} = \gamma_{si} I(\gamma_{si} > \rho)$ with some pre-set threshold ρ.

As a default choice for δ_{sim} [22] proposes for continuous covariates to use a rescaled version of the Mahalanobis distance. Let $d_{si} = d_{MH}(D_{si}, D_0)$ using sample mean and covariance matrix of D_0 to evaluate the Mahalanobis distance, and define $\gamma_{si} = 1 - r \cdot d_{si}$, where r is chosen to map d_{si} to the unit interval. For other data formats, Golchi [22] suggests to use the posterior predictive distribution for D_{si} given D_0, i.e. $d_{si} = p(D_{si} \mid D_0) = \int p(D_{si} \mid \psi) \cdot \pi_0(\psi) \prod_j p(D_{0j} \mid \psi)\, d\psi$ using the same probability models as the inference model, with ψ replacing θ. The expression is assuming that the inference model (or a simplified version of it) takes the form of i.i.d. sampling given θ.

3. A common-atoms BNP mixture model for using RWD

Chandra *et al.* [2] develop a model-based approach to include RWD in clinical trial design using nonparametric Bayesian (BNP) mixture models. We introduce the set up and approach in the context of a specific example.

Example 3.1. Synthetic controls for single-arm glioblastoma (GBM) trials. For a variety of reasons (e.g. [3]) many phase II trials in GBM are designed as single-arm treatment-only studies (SAT). Let then $s = 0$ indicate a current SAT, and assume that RWD is available in the form of data for earlier patients in the same hospital and treated by the same clinical group ($s = 1$). We want to use this RWD to create a synthetic control arm to augment the current SAT study. See [2] for more details. A summary of the data is shown in figure 1. We report results for this application in the next section, after introducing the general approach. In the following discussion we will use (T) and (C) to refer to the patient population in the SAT ('T' for treatment), and the RWD ('C' for control), respectively.

(a) Common atoms mixtures

The approach proceeds as follows. First we fit a BNP mixture model, $F_1(x_{1i})$, to patient covariates x_{1i} in (C). For simplicity, assume a mixture of normals. Then we fit a second mixture model, $F_0(x_{0i})$, to (T), using the same atoms, that is, the same normal location-scale parameters in the case of the mixture of normals. Only the weights distinguish F_1 and F_0. Let π_{sh}, $s = 0, 1$, denote these weights. Figure 2 illustrates the two mixture models F_0 and F_1, assuming univariate x_{si} and four components, $h = 1, \ldots, 4$. Part of the inference is a latent allocation of all data points, i.e. patients, to one of the terms in the mixture. This link is not fixed. It involves random indicator variables c_{si} that link patient i in study s with one of the terms in the mixture. We are already almost done. Clearly the desired adjustment of the RWD data to create an equivalent patient population should involve the ratio of the weights in the mixture. For patient i in (C), let then w_i denote the posterior expectation of the ratio $\pi_{0c_{1i}}/\pi_{1c_{1i}}$. It can be shown that thinning out the RWD by retaining data points with probability w_i creates an adjusted distribution $F_1' = F_0$. One could then proceed with the desired inference as if the study were a randomized controlled trial. Figure 3 illustrates the idea. For this scheme it is critical that the nonparametric mixture model F_s can fit the data. Under appropriate (mostly technical) assumptions this is the case. BNP models 'are always right', in the sense of full prior support (e.g. [23]). We skipped some details in this brief outline. In particular, the atoms of F_0 are only defined conditional on latent cluster allocations for the real-world data. See below for more details.

For a formal description of the same scheme let $x_{si} = (x_{sij}, j = 1, \ldots, p)$ denote the baseline covariates for patient i in study s, including p covariates. Fitting a mixture model to x_{1i}, we assume

$$x_{1i} \sim \underbrace{\int q(x_{1i}; \zeta)\, dG_1(\zeta)}_{F_1(x)}, \quad i = 1, \ldots, n_1, \tag{3.1}$$

where G_1 is a random mixing measure. Let δ_x denote a point mass at x. Using a prior probability model that restricts G_1 to discrete probability measures we can represent G_1 as $G_1 =$

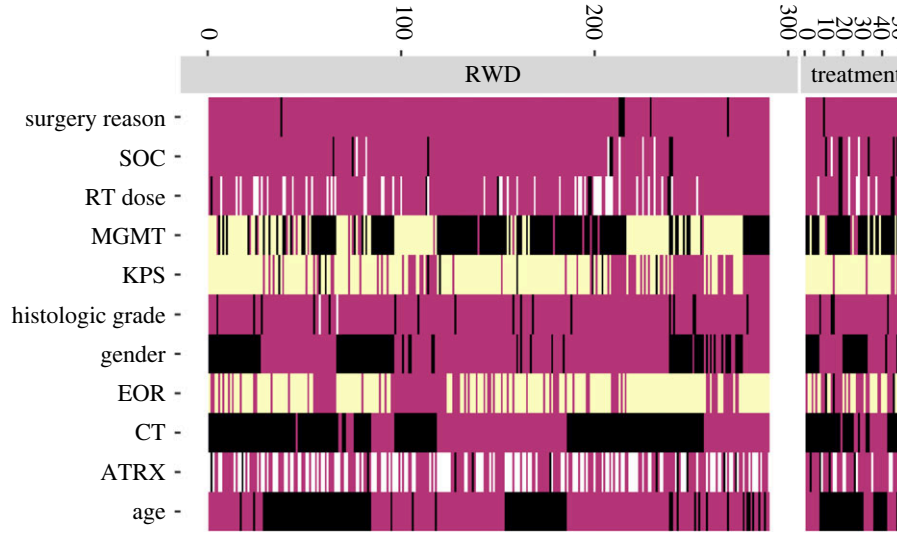

Figure 1. Data for historical data (RWD) (left block) and a current single arm, treatment-only study (right block). The horizontal axis are patients, and the vertical axis are $p = 11$ baseline covariates. All covariates are categorical, with different colours indicating different levels of each covariate (0 = black, 1 = purple or dark grey, 2 = yellow or light grey) and white for missing data. (Online version in colour.)

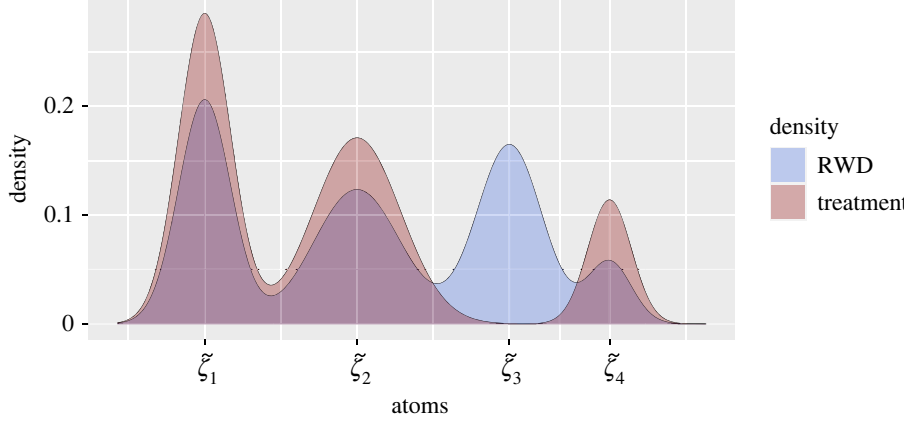

Figure 2. Two common-atoms mixture models for the RWD (blue, with four terms) and the SAT (red, with three terms). The mixture models share the same kernels $q(\cdot; \tilde{\zeta}_h)$ and differ only by the weights π_{sh}. (Online version in colour.)

$\sum_h \pi_{1h}\, \delta_{\tilde{\zeta}_h}$. Using, for example, multivariate normal kernels $q(x; \zeta = (\mu, \Sigma)) = N(x; \mu, \Sigma)$, equation (3.1) becomes a mixture of normal model. The latter can be written as

$$F_1 = \sum_h \pi_{1h} q(\bullet; \tilde{\zeta}_h) \quad \text{or} \quad \begin{cases} x_{1i} \mid \zeta_i \sim q(\cdot; \zeta_i) \\ \zeta_i \sim G_1. \end{cases} \tag{3.2}$$

Sampling from the mixture induces a random partition of $[n_1] = \{1, \ldots, n_1\}$. First, note that the mixture model can equivalently be written as $x_{1i} \mid \zeta_i \sim q(\cdot; \zeta_i)$ with $\zeta_i \sim G_1$, as indicated on the right side of (3.2). The discrete nature of G_1 implies a positive probability of ties among the ζ_i. The arrangement of such ties defines clusters C_{1k}. Let $\{\zeta_1^\star, \ldots, \zeta_K^\star\}$ denote the $K \leq n_1$ unique values among the ζ_i. Finally, let $c_{1i} = k$ when $i \in C_{1k}$, i.e. $\zeta_i = \zeta_k^\star$. We will use these cluster membership

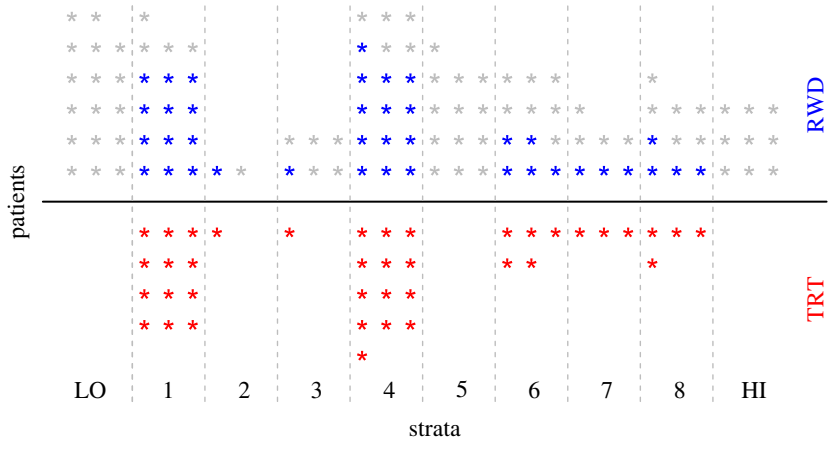

Figure 3. Matching the two populations by thinning out. Blue stars (top) indicate patients in the external data, arranged by strata. Red stars (bottom) are patients in the treatment arm. Stars in light grey (top) indicate patients in the external data that are (conceptually) dropped from the dataset to create equivalent populations. (Online version in colour.)

indicators c_{1i} to represent the clusters in posterior simulation below. The described structure is summarized as a partition $\rho_1 \equiv \{C_{11}, \ldots, C_{1K}\}$ of $[n_1]$ into clusters C_{1k}. That is, $C_{1k} \subseteq [n_1]$ and $\dot{\bigcup}_k C_{1k} = [n_1]$. Here $\dot{\bigcup}$ refers to a disjoint union with $C_{1k} \cap C_{1\ell} = \emptyset$ for $k \neq \ell$. See, for example, [24] for more discussion of this random partition.

The idea is then to fit the mixture F_1 to the RWD, generate the random partition $\{C_{11}, \ldots, C_{1K}\}$ and cluster-specific parameters $\zeta_k^\star, k = 1, \ldots, K$. Note that the random number of clusters, K, is part of the random partition. Also, for easier notation assume without loss of generality that $\zeta_k^\star = \tilde{\zeta}_k$, $k = 1, \ldots, K$ (re-indexing $\tilde{\zeta}_h$ if needed). Given $\zeta^\star$ we then fit a second mixture

$$F_0 = \sum_{k=1}^{K} \pi_{0k} q(\bullet; \zeta_k^\star), \tag{3.3}$$

to the SAT data $x_{0i}, i = 1, \ldots, n_0$. Two observations about F_0 are in order. First, F_0 is a mixture with respect to a random probability measure with atoms matching those of F_1 only that are linked to data points under c_{1i}. Second, in F_s, $s = 0, 1$ only the weights π have s indices. The atoms $\zeta^\star$ are shared (more precisely, the subset of those atoms that are allocated under c_{1i} are shared). Such 'common atoms' BNP mixture models have been shown to be useful in other applications too [25,26].

In summary, the common atoms mixture model (CAM) for RWD and the current study is as follows:

$$\left.\begin{aligned} p(x_{1i} \mid \tilde{\zeta}, \pi_1) &\sim \overbrace{\sum_h \pi_{1h} q(x_{1i} \mid \tilde{\zeta}_h)}^{F_1(x_{1i})} \text{ or } \begin{cases} p(x_{1i} \mid \zeta_{1i}) & \sim \quad q(x_{1i} \mid \zeta_{1i}) \\ p(\zeta_{1i} = \tilde{\zeta}_h) & = \quad \pi_{1h} \end{cases} \\ p(x_{0i} \mid \zeta^\star, \pi_0) &\sim \underbrace{\sum_k \pi_{0k} q(x_{0i} \mid \zeta_k^\star)}_{F_0}. \end{aligned}\right\} \tag{3.4}$$

Recall that $\zeta^\star$ is defined as the set of unique values in $\{\zeta_{1i},\ i = 1, \ldots, n_1\}$, and we assume $\tilde{\zeta}_k = \zeta_k^\star$ (re-arranging indices if needed). In other words, the prior on F_0 is defined conditional on F_1 and $\zeta^\star$. For reasons explained later, in §3c, we refer to model (3.4) as CA-PPMx. The model is illustrated in figure 2. Similar common-atoms mixture models, without the constraint on the second mixture model to only use the atoms $\zeta_k^\star$ sampled in the first mixture, are used, for example, in [25]. The

latter constraint makes the mixture F_0 dependent on the latent $\zeta^\star$, which are therefore introduced in the first line of (3.4).

We implement MCMC-based posterior simulation under this model, obtaining a posterior Monte Carlo sample of $\pi_{1k}, \pi_{0k}, c_{1i}$. Let $m_{1k} = |C_{1k}|$ denote the size of the kth cluster under the RWD. Let $D_s = \{x_{si},\ i = 1, \ldots, n_s\}$ denote the data. We use the posterior Monte Carlo sample to evaluate the posterior mean $w_i \propto E(\pi_{0c_{1i}}/m_{1c_{1i}} \mid D_0, D_1)$, $i = 1, \ldots, n_1$. Here we replaced π_{1k} in the denominator by the cluster size (to avoid numerical problems). We use these weights to thin out the RWD sample, to create an adjusted sample from $F_1' = F_0$, as desired.

(b) Algorithm

The approach is summarized by the following four steps. Let (T) and (C) denote the single-arm treatment only study (SAT) and RWD data, respectively.

STEP 1: Fit data in (T) and (C) , using a *common atoms random partition* (3.4) including $F_1, \xi^\star$ and F_0.
STEP 2: Resample (-weigh) the (C) patients, to achieve equivalent patient populations.
STEP 3: Prove equivalence, using any general purpose classification (support vector machine, Bayesian random forests, etc.) of the merged data set, merging (T) and (C). Failure to classify the merged data proves equivalence.

At this moment, we have (T) $\approx$ resampled/weighted (C); and could proceed as if the resampled/weighted (C) were a control arm in a RCT.

STEP 4: Carry out inference on treatment effects as if the study were an RCT.
STEP 4′: Alternatively, extend the BNP mixture model (3.4) for x_{si} to include outcomes y_{si}, allowing model-based inference on treatment effects.

The latter allows us to impute potential outcomes and proceed by reporting a model-based average treatment effect as a posterior expected difference of outcomes under treatment versus control.

Some more details on STEP 2: from (3.4), it is intuitively plausible that reweighing of the control patients to evaluate expectations under F_0 should use the already mentioned weights $w_i = \pi_{0k}/\pi_{1k}$, with $k = c_{1i}$. More specifically, in [2] we show that expectations under F_0 can be evaluated using posterior Monte Carlo samples conditional on D_1 after thinning out the external data using weights w_i. We will use this in STEP 4′ to evaluate posterior expectations of average treatment effects. See [2] for a formal discussion and justification. In the implementation, we evaluate the weights for $i = 1, \ldots, n_1$ as a Monte Carlo average

$$w_i \propto \sum_t \frac{\pi_{0c_{1i}}}{m_{1c_{1i}}/n_1},$$

where t indexes posterior Monte Carlo samples, $m_{1k} = |C_{1k}|$ is the size of cluster k, and c_i is the cluster membership indicator with $c_i = k$ when $i \in C_k$.

Finally, more detail on the evaluation of treatment effects in STEP 4. Note that the model in (3.4) is restricted to patient covariates x_{si}. It does not yet include a sampling model for the outcomes. But the model can easily be extended to include outcomes y_{si} by introducing cluster-specific parameters θ_{sk} and assuming

$$y_{si} \sim p(y_{si} \mid i \in C_{sk}, \theta_{sk}),$$

where θ_{0k} and θ_{1k}, $k = 1, \ldots, K$, index the response model under treatment and control, respectively. We then define a cluster-specific treatment effect using a suitable function $d(\theta_{0k}, \theta_{1k})$ of these parameters. For example, if the interpretation of θ_{sk} is as an average event time, this

could be $d_k = (\theta_{0k} - \theta_{1k})$. Finally, we report the overall treatment effect under F_0 as the posterior expectation of

$$\delta = \sum_k \pi_{0k} d(\theta_{0k}, \theta_{1k}), \tag{3.5}$$

with the posterior expectation averaging over the random partition, the weights and the treatment effect parameters.

(c) Implied random partition and missing data

Some more observations on $\rho_1 = \{C_{11}, \ldots, C_{1K}\}$. Since we only discuss the partition of (C) here, we for the moment drop the study subindex $_1$. The discrete mixture (3.1) for the external data implies the already mentioned random partition of patient indices $[n] \equiv \{1, \ldots, n\}$ into clusters $\rho = \{C_1, \ldots, C_K\}$. A widely used choice for the prior probability model for G_1 in (3.1) is the Dirichlet process (DP) prior [27]. The DP prior on G_1 implies a marginal prior $p(\rho)$ which is known as the Chinese restaurant process and can be written as $p(\rho) \propto \prod_k (|C_k| - 1)!$ [28]. In general, a random partition of this product form is known as product partition model (PPM) [29]. The factors $c(C_k)$, in this case, $c(C_k) = (|C_k| - 1)!$ are known as the concentration function. Let $x^\star_k = \{x_i;\ i \in C_k\}$ denote the covariate vectors arranged by cluster. Conditional on x the model becomes

$$p(\rho \mid x) \propto \prod_k c(C_k)\, g(x^\star_k). \tag{3.6}$$

Here, the last factor is $g(x^\star_k) = \int \prod_{i \in C_k} q(x_i \mid \zeta^\star_k)\, \mathrm{d}G^\star(\zeta^\star_k)$, with $q(\cdot)$ referring to the mixture kernel and $G^\star$ is the prior on the cluster-specific parameters $\zeta^\star_k$, which is specified as one of the hyperparameters of the DP prior on G_1. The random partition $p(\rho \mid x)$ again takes the form of a PPM, now indexed with covariates x_i. In other words, $p(\rho \mid x)$ defines a regression of the random partition ρ on covariates x [30]. In the context of this model, the additional factor $g(x^\star_k)$ serves a purpose like a purity or 'similarity' function in hierarchical clustering algorithms. It favours the formation of clusters with similar x_i. For more discussion of such models for prediction, see also [31]. We refer to (3.6) as the PPMx model, and to (3.4) as the CA-PPMx model to highlight the implied random partition.

(i) Missing data

An important feature of the model for the application to RWD is the natural accommodation of missing data (missing at random). Assuming that the kernels factor across coordinates of x_i as $q(x_i \mid \zeta^\star_k) = \prod_j q_j(x_{ij} \mid \zeta^\star_{kj})$, the similarity function $g(x^\star_k)$ can be evaluated using available covariates x_{ij} only. For each patient, let $\mathcal{C}_i \subset \{1, \ldots, p\}$ denote the indices of all available covariates, let $x_i = (x_{ij},\ j \in \mathcal{C}_i)$ and let $x^\star_k = \{x_i;\ i \in C_k\}$ denote the set of only the observed covariates for all patients in cluster C_k. Also, let $\mathcal{O}_{kj} = \{i :\ i \in C_k \text{ and } j \in \mathcal{C}_i\}$. Then

$$g(x^\star_k) = \int \prod_{i \in C_k} q(x_i \mid \zeta^\star_k)\, dG^\star(\zeta^\star_k) = \prod_{j=1}^{p} \int \prod_{i \in \mathcal{O}_{kj}} q(x_{ij} \mid \zeta^\star_{kj})\, dG^\star(\zeta^\star_{kj}). \tag{3.7}$$

In the last expression, the product over i goes only over all those patients in cluster C_k with available jth coordinate. Missing covariate values are simply not used.

Implicit in this construction is the assumption of missing at random. An interesting situation could arise when RWD includes multiple sources of datasets with different sets of missing covariates. While formally (3.7) can accommodate this set-up, one would need to carefully consider whether it is reasonable to continue assuming missing at random. Missingness patterns other than completely at random can be handled by introducing additional hierarchy in the model (see, e.g. [32] for a review).

4. A Glioblastoma study

In this section, we discuss an application example and a simulation study to illustrate the use of RWD for creating a synthetic control arm to complement a single-arm treatment-only study.

Example 4.1. (continued): Synthetic controls for single-arm glioblastoma (GBM) trials. In [2] we apply the CAM approach to create a synthetic control arm to design a single arm treatment-only phase II trial for GBM patients treated at M.D. Anderson Cancer Center. We use data from historical patients treated over recent years in the same clinic. This RWD includes $n_1 = 339$ patients. We record $p = 11$ carefully selected covariates. Given the devastating nature of glioblastoma thorough studies have been conducted to find clinically relevant covariates [33,34]. We followed these earlier studies and expert advice from clinical collaborators to choose the selected covariates.

Consider then a new SAT with $n_0 = 49$ patients. The endpoint is overall survival (OS). Let λ_t denote the hazard ratio at time t for patients under treatment versus control. The evaluation of the hazard ratio averages over the covariates x_i of a future patient $i = n + 1$ and is evaluated over a range of t from 0 to 180 weeks. A new treatment is considered clinically effective if $\lambda_t < 0.6$ for all t.

In the design, we consider two scenarios. In the first scenario (null scenario, H_0), we assume $\lambda_t = 1$, that is, no difference between treatment and control. In the second scenario (alternative scenario, H_1), we assume a treatment effect with $\lambda_t < 0.6$ for $0 < t < 180$. Figure 4 shows Kaplan–Meier (KM) plots for one hypothetical realization of the trial under H_0 (*a*) and under H_1 (*b*).

For a more extensive comparison, we carry out a simulation study. We compare the CA-PPMx model approach with an approach using a PS-integrated power prior (PP) [17], and a similar, non-Bayesian propensity-score composite likelihood approach (CL) [18] (compare with the discussion of the PP and CL approach in §2c).

We simulate hypothetical data under three scenarios. We generate covariates x_i, $i = 1, \ldots, n$, from a model $p(x)$ for the (assumed) marginal distribution of patient covariates in a merged population. Here $n = n_0 + n_1$ is the total number of patients (and see next for n_0 and n_1). Given x_i we then select each patient into either the external data ($S_i = 1$) or the treatment only single arm trial ($S_i = 0$), with $n_1 = \sum S_i$ and $n_0 = n - n_1$. That is, instead of simulating x_{si} for $s = 0, 1$ separately we first simulated x_i for the merged population, and then only assigned patients to one or the other study.

For $p(x)$, we used the empirical distribution of the data in the historical GBM studies. To create a desired simulation truth δ^0 for the treatment effect, we use the recorded outcomes for patients under $S_i = 1$ and increment outcomes for patients under $S_i = 0$. To select patients into one or the other studies, we used three different (simulation truth) propensity score models: (i) a logistic regression without interaction, as assumed by the PP and CL methods. We therefore refer to this as the *Oracle* scenario; (ii) a logistic regression with interaction (*Interaction* scenario); and (iii) using the propensity score model implied by the CA mixture model (*CA Mixture*). For the simulation under scenarios (i) and (ii), we use $n_0 = 49, n_1 = 290$ and $p = 11$ with all categorical covariates; and for scenario (iii) $n_0 \in \{50, 100, 150\}$, $n_1 = 6 \cdot n_0$ and $p \in \{10, 15, 20\}$. The scenarios were chosen to mimic some aspects of the GBM study. In fact, the covariate distribution $p(x)$ for (i) and (ii) is defined as the empirical distribution of the $n_1 = 339$ historical patients in that study.

Under each hypothetical trial realization, we estimated the treatment effect δ. We define a simulation truth of the treatment effect as $\delta^o \equiv (1/n) \sum E\{Y_i(1) - Y_i(0)\}$ as average difference of mean outcome under treatment and control for patients $i = 1, \ldots, n$, with the expectation evaluated under the simulation truth. Let then $bias = |\delta^0 - E(\delta \mid y)|$ for true treatment effect δ^0 denote the bias of the reported treatment effect. Figures 5 and 6 summarize results under the three scenarios. The figures show boxplots of realized bias under 100 repeat simulations, arranged by sample size n_1, number of covariates p, true treatment effect δ^o and method. For each combination of (n_1, p, δ^o), the figures show five boxplots corresponding to inference under the CA mixture model, CL and PP, using logistic and random forest propensity score models for the latter two.

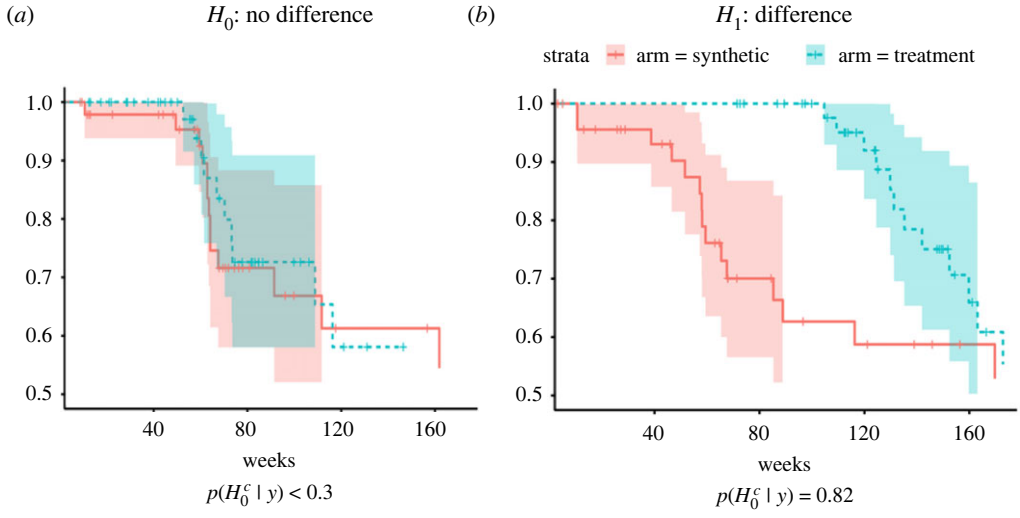

Figure 4. KM plots for one hypothetical realization of the trial using a synthetic treatment cohort under H_0 (*a*) and H_1 (*b*). (Online version in colour.)

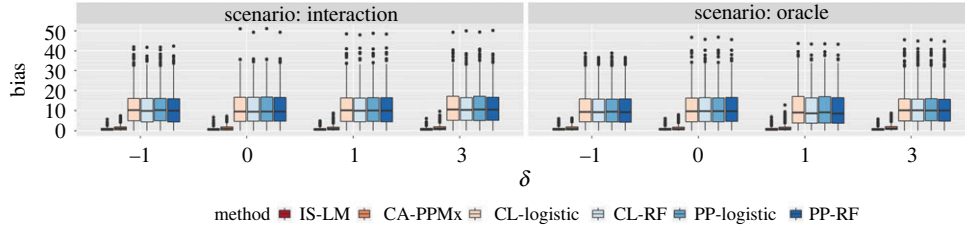

Figure 5. Summaries of simulation results under the *oracle* scenario (i) (right panel), and the *interaction* scenarios (ii) (left panel). The boxplots show realized bias over 100 repeat simulations under assumed true treatment effects $\delta \in \{-1, 0, 1, 3\}$, and under each of the compared methods. See the text for an explanation of the six compared methods. (Online version in colour.)

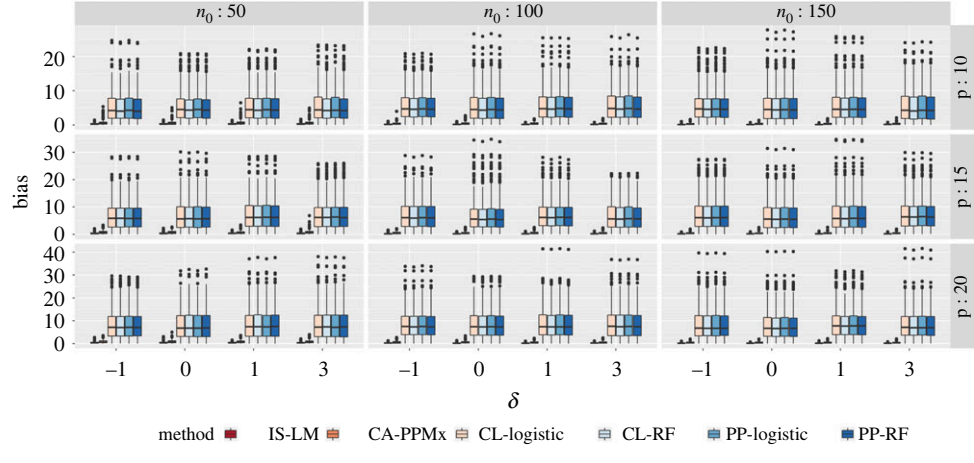

Figure 6. Summaries of simulation results under the *CAM* scenario (iii) showing realized bias over 100 repeat simulations under assumed true treatment effects $\delta \in \{-1, 0, 1, 3\}$, and under each of the compared methods. See the text for the six compared methods. (Online version in colour.)

The labels for the compared methods are as follows. CA-PPMx refers to model-based inference using the proposed CA mixture model and (3.5), that is, using steps 1,2,3 and 4 of the algorithm in §3c; IS-LM refers to inference using steps 1,2,3 and 4 of the algorithm, with a linear model to compare treatment effects in step 4; PP-Logistic, PP-RF, CL-Logistic and CL-RF are PP and CL approaches using a logistic regression and random forest [35] PS model, respectively. The latter four methods are briefly discussed in §2c.

Overall the simulation results are favourable for the CAM model. The increased bias under the PS-based methods might be due to the multi-step nature of the methods, ignoring substantial uncertainty in the propensity scores $e(x)$, the realistic, but small sample sizes relative to the $p \geq 10$ covariates, and model misspecification in scenarios (ii) and (iii).

5. Conclusion

We have briefly discussed several methods for including RWD in clinical studies, focusing mainly on Bayesian methods. The main challenge is to adjust for the lack of randomization, to coherently propagate all uncertainties, and to combine different sources of information. All three considerations are naturally formalized and can be addressed under a Bayesian framework. There are no magic solutions. But a principled Bayesian approach can clearly articulate all assumptions, quantify uncertainties and allow us to systematically study sensitivity and error rates.

One limitation of the methods that we discussed here is the dimension of the data. In many cases, when data collection is automated and/or involves large scale genomic markers the dimension of patient-specific covariates can easily exceed what is computationally feasible for inference with nonparametric mixture models. Matching subpopulations with respect to all reported covariates becomes increasingly more restrictive with increasing dimension covariate vectors. Also, achieving matching distributions on all available covariates becomes increasingly less important as additional covariates add progressively more noise unrelated to treatment effects. One would need to include dimension reduction as part of the inference pipeline, using, for example, methods proposed in [36] or [37]. Another limitation when working with multiple data sets is the coherent definition of variables and coding. This is a critical precondition to any meaningful inference. See [38] for a recent discussion.

Finally, throughout we proceeded assuming that all relevant covariates are recorded, that is, there are no unmeasured confounders. For carefully planned clinical studies, this assumption is reasonable for most instances of statistical inference. For example, in the case of the GBM study, there is a wide consensus about the set of relevant baseline covariates. We would still recommend to follow up with a sensitivity analysis, perhaps in the form of a simulation study to investigate the size of plausible effects of unmeasured confounders.

Data accessibility. This article has no additional data.

Authors' contributions. P.M.: conceptualization, formal analysis, investigation, methodology, project administration, writing—original draft, writing—review and editing; N.C.: conceptualization, formal analysis, investigation, methodology, writing—original draft, writing—review and editing; A.S.: conceptualization, investigation, methodology, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. No funding has been received for this article.

References

1. FDA. 2018 *Framework for FDA's real-world evidence program*. Food and Drug Administration.
2. Chandra NK, Sarkar A, de Groot J, Yuan Y, Müller P. 2022 Bayesian nonparametric common atoms regression for generating synthetic controls in clinical trials. (http://arxiv.org/abs/2201.00068).

3. Grossman SA, Schreck KC, Ballman K, Alexander B. 2017 Point/counterpoint: randomized versus single-arm phase II clinical trials for patients with newly diagnosed glioblastoma. *Neuro Oncol.* **19**, 469–474. (doi:10.1093/neuonc/nox030)
4. Ventz S, Comment L, Louv B, Rahman R, Wen PY, Alexander BM, Trippa L. 2021 The use of external control data for predictions and futility interim analyses in clinical trials. *Neuro Oncol.* **24**, 247–256. (doi:10.1093/neuonc/noab141)
5. Gelfand AE, Smith AFM. 1990 Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* **85**, 398–409. (doi:10.1080/01621459.1990.10476213)
6. Smith AFM, Roberts GO. 1993 Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Stat. Soc.: Ser. B (Methodological)* **55**, 3–23.
7. Walker SG, Damien P, Laud PW, Smith AFM. 1999 Bayesian nonparametric inference for random distributions and related functions. *J. R. Stat. Soc.: Ser. B (Statistical Methodology)* **61**, 485–527. (doi:10.1111/1467-9868.00190)
8. Denison DG, Mallick BK, Smith AF. 1998 A Bayesian CART algorithm. *Biometrika* **85**, 363–377.
9. Bernardo JM, Smith AFM. 2009 *Bayesian theory*, **vol. 405**. New York, NY: John Wiley & Sons.
10. Chen MH, Ibrahim JG. 2000 Power prior distributions for regression models. *Stat. Sci.* **15**, 46–60. (doi:10.1214/ss/1009212673)
11. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. 2011 Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* **67**, 1047–1056. (doi:10.1111/j.1541-0420.2011.01564.x)
12. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. 2010 Summarizing historical information on controls in clinical trials. *Clin. Trials* **7**, 5–18. (doi:10.1177/1740774509356002)
13. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. 2014 Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* **70**, 1023–1032. (doi:10.1111/biom.12242)
14. Rosenbaum PR, Rubin DB. 1983 The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55. (doi:10.1093/biomet/70.1.41)
15. Zhou T, Ji Y. 2021 Incorporating external data into the analysis of clinical trials via Bayesian additive regression trees. *Stat. Med.* **40**, 6421–6442. (doi:10.1002/sim.9191)
16. Chipman HA, George EI, McCulloch RE. 2010 BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4**, 266–298. (doi:10.1214/09-AOAS285)
17. Wang C, Li H, Chen WC, Lu N, Tiwari R, Xu Y, Yue LQ. 2019 Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *J. Biopharm. Stat.* **29**, 731–748. (doi:10.1080/10543406.2019.1657133)
18. Chen WC, Wang C, Li H, Lu N, Tiwari R, Xu Y, Yue LQ. 2020 Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. *J. Biopharm. Stat.* **30**, 508–520. (doi:10.1080/10543406.2020.1730877)
19. Wang C. 2022 psrwe: PS-integrated methods for incorporating RWE in clinical studies. R package, version 3.1, https://github.com/olssol/psrwe.
20. Liu M, Bunn V, Hupf B, Lin J, Lin J. 2021 Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data. *Stat. Med.* **40**, 4794–4808. (doi:10.1002/sim.9095)
21. Wang C, Rosner GL. 2019 A Bayesian nonparametric causal inference model for synthesizing randomized clinical trial and real-world evidence. *Stat. Med.* **38**, 2573–2588. (doi:10.1002/sim.8134)
22. Golchi S. 2020 Use of historical individual patient data in analysis of clinical trials. (http://arxiv.org/abs/2002.09910).
23. Ghosal S, van der Vaart A. 2017 *Fundamentals of nonparametric Bayesian inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK: Cambridge University Press.
24. Müller P. 2018 Bayesian nonparametric mixture models. In *Handbook of mixture analysis* (eds S Frühwirth-Schnatter, G Celeux, C Robert), ch. 6, pp. 101–121. Boca Raton, FL: Chapman and Hall/CRC.
25. Denti F, Camerlenghi F, Guindani M, Mira A. 2021 A common atoms model for the Bayesian nonparametric analysis of nested data. *J. Am. Stat. Assoc.* (doi:10.1080/01621459.2021.1933499)

26. Sarkar A. 2022 Bayesian semiparametric covariate informed multivariate density deconvolution. *J. Comput. Graph. Stat.* **31**, 20222060239. (doi:10.1080/10618600.2022.2060239)
27. Ferguson TS. 1973 A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230. (doi:10.1214/aos/1176342360)
28. Müller P, Quintana FA, Jara A, Hanson T. 2015 *Bayesian nonparametric data analysis*, vol. 1. Berlin, Germany: Springer.
29. Hartigan JA. 1990 Partition models. *Commun. Stat.-Theory Methods* **19**, 2745–2756. (doi:10.1080/03610929008830345)
30. Müller P, Quintana F, Rosner GL. 2011 A product partition model with regression on covariates. *J. Comput. Graph. Stat.* **20**, 260–278.
31. Page GL, Quintana FA, Müller P. 2022 Clustering and prediction with variable dimension covariates. *J. Comput. Graph. Stat.* **31**, 466–476. (doi:10.1080/10618600.2021.1999824)
32. Linero AR, Daniels MJ. 2018 Bayesian approaches for missing not at random outcome data: the role of identifying restrictions. *Stat. Sci.* **33**, 198–213. (doi:10.1214/17-STS630)
33. Nam JY, de Groot JF. 2017 Treatment of glioblastoma. *J. Oncol. Practice* **13**, 629–638. (doi:10.1200/JOP.2017.025536)
34. Alexander BM et al. 2019 Individualized screening trial of innovative Glioblastoma therapy (INSIGhT): a Bayesian adaptive platform trial to develop precision medicines for patients with Glioblastoma. *JCO Precision Oncol.* **3**, 1–13.
35. Breiman L. 2001 Random forests. *Mach. Learn.* **45**, 5–32. (doi:10.1023/A:1010933404324)
36. Chandra NK, Canale A, Dunson DB. 2021 Escaping the curse of dimensionality in Bayesian model-based clustering. (http://arxiv.org/abs/2006.02700).
37. De Vito R, Bellio R, Trippa L, Parmigiani G. 2019 Multi-study factor analysis. *Biometrics* **75**, 337–346. (doi:10.1111/biom.12974)
38. Meng XL. 2021 Enhancing (publications on) data quality: deeper data minding and fuller data confession. *J. R. Stat. Soc.: Ser. A (Statistics in Society)* **184**, 1161–1175. (doi:10.1111/rssa.12762)