A Two-Stage Bayesian Framework for Combining Clinical Trial and Real-World Data: Model-Based Clustering and Hierarchical Modeling

Hyotae Kim Susan Halabi

Department of Biostatistics & Bioinformatics, Duke University
hyotae.kim@duke.edu susan.halabi@duke.edu

Abstract

Real-world data (RWD) can complement randomized clinical trial (RCT) data by providing larger and more diverse patient populations. However, integrating RWD with RCT data is complicated by differences in baseline characteristics and outcomes across data sources. We propose a novel two-stage Bayesian framework for integrative analyses of RCT and RWD. In the first stage, a model-based clustering approach promotes baseline comparability by grouping patients with similar observed covariate distributions. The clustering model incorporates a newly developed quasi-geometric weights (QGW) prior that provides greater flexibility than the geometric weights prior while preserving ordered mixture weights. In the second stage, a Bayesian hierarchical model estimates treatment contrasts while accommodating heterogeneity across clusters and data sources and adaptively regulating information borrowing from RWD. We further formulate restricted mean survival time-based causal estimands within the potential outcomes framework and extend the model to single-arm trials with external control data. Simulation studies evaluate clustering recovery, treatment-contrast estimation under varying degrees of control-group scarcity and clustering accuracy, performance under systematic RCT-RWD discrepancies, and comparisons with a propensity score-integrated power prior approach. We apply the proposed framework to a prostate cancer trial, integrating multiple external data sources in a single-arm setting, with exploratory results suggesting worse survival in two patient subgroups.

Keywords: Bayesian model-based clustering; Bayesian hierarchical model; Geometric weights prior; Randomized clinical trial; Real-world data.

1 Introduction

There has been growing interest in using real-world data (RWD) to support healthcare decision-making. Initiatives such as the 21st Century Cures Act and recent FDA guidances (e.g., "Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices"; "Framework for FDA's Real-World Evidence Program"; "Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drug and Biological Products") reflect this shift toward integrating RWD into regulatory processes. RWD are derived from sources such as electronic health records, insurance claims databases, and disease registries and may provide large and diverse patient populations at relatively low cost (Liu and Panagiotakos, 2022). In contrast, randomized clinical trials (RCTs), often considered the gold standard for treatment-effect estimation, are conducted under carefully

controlled conditions, with random treatment assignment helping to reduce confounding and bias. However, RCTs are typically expensive and time-consuming and may enroll relatively small numbers of patients because of restrictive eligibility criteria, particularly in settings such as rare diseases. These limitations have motivated increasing interest in supplementing RCT data with external information.

As RWD can serve as an effective supplement to RCT data, analyses exploiting combined RCT and RWD sources have been actively explored (e.g., Lin et al., 2022; Tang et al., 2023; Vaghela et al., 2024). However, such data integration poses several well-known challenges, typically summarized into two main categories: (1) *baseline incomparability* (also referred to as *covariate shift*), and (2) *outcome incomparability* (also known as *outcome drift*). The former refers to differences in baseline characteristics between patients in RWD and RCTs, leading to shifts in the distribution of baseline covariates. The latter arises when outcomes in RWD differ from those in RCT due to factors unrelated to treatment, such as unmeasured confounding, lack of concurrency, measurement error, and/or misalignment in time zero or outcome assessment.

For baseline comparability, matching and weighting methods (e.g., Shan et al., 2022; Hainmueller, 2012; Lin et al., 2019) are widely used as a preprocessing step before outcome modeling. Propensity score matching, in particular, has garnered considerable attention and is adopted across a wide range of modeling frameworks (e.g., Wang et al., 2019, 2020; Chen et al., 2020).

Following covariate adjustment, integrative outcome analyses are frequently conducted using Bayesian dynamic borrowing methods. The power prior, introduced by Ibrahim and Chen (2000), incorporates the likelihood contribution from the RWD, down-weighted by a power parameter, into the prior distribution for the RCT model parameters. Hobbs et al. (2012) proposed the commensurate prior, which centers the prior distribution for the RCT parameters on estimates from the external data and uses a commensurability parameter to regulate the degree of borrowing. Power priors provide a simple and intuitive mechanism for discounting external information, whereas commensurate priors allow borrowing to adapt to the degree of compatibility between the current and external data through a random commensurability parameter. Jiang et al. (2023) developed the elastic prior, which regulates borrowing through an elastic function of a data-driven measure of discrepancy between the two data sources. By explicitly linking a congruence measure to the strength of borrowing through a calibrated elastic function, this approach provides more direct control over borrowing and facilitates type I error rate control. Furthermore, Schmidli et al. (2014) and Yang et al. (2023) developed mixture priors that combine informative and noninformative components, such as a meta-analytic predictive (MAP) prior and a vague prior. These approaches provide a natural framework for synthesizing information across multiple historical studies while protecting against excessive borrowing: under substantial prior-data conflict, the borrowing weight shifts toward the vague component, reducing the influence of historical information when the current and external data disagree.

Despite these methodological developments, a key challenge remains: covariate comparability between RCT and external-data populations may vary substantially across different regions of the covariate space. A global assessment of covariate balance or overlap may therefore obscure important subgroup-specific discrepancies and mask heterogeneity in the baseline covariate distributions. For example, Wang et al. (2019) partitions patients into propensity-score strata and demonstrates that the distributions of several covariates differ across strata (Figure 3), subsequently assigning stratum-specific power-prior weights according to the degree of overlap between the RCT and external-data populations. Nevertheless, many existing borrowing approaches rely on global model

parameters and a common borrowing mechanism and therefore do not explicitly accommodate heterogeneity in both covariate comparability and outcome relationships across patient subgroups. This motivates a framework that identifies subgroups of patients with similar baseline characteristics and allows treatment contrasts and cross-source discrepancies to vary across these subgroups.

We propose a novel two-stage Bayesian framework combining clustering and hierarchical modeling to jointly analyze RCT and RWD. In the first stage, we develop a Bayesian model-based clustering approach using a new nonparametric quasi-geometric weights (QGW) prior, which provides a flexible mixture representation while retaining an ordered-weight structure that mitigates label switching and allows the number and composition of latent clusters to be learned from the data. In the second stage, we develop a Bayesian hierarchical model incorporating cluster- and study-specific effects to account for heterogeneity across latent patient subgroups and data sources, while providing cluster-specific and overall treatment-effect estimates. RWD-specific discrepancy parameters further account for residual outcome differences between RCT and RWD and adaptively regulate the contribution of RWD according to cross-source compatibility, although bias due to unmeasured confounding cannot be fully eliminated through modeling alone.

The remainder of the paper is organized as follows. Section 2 provides background on the geometric weights prior. Section 3 introduces the proposed two-stage framework. Section 3.1 presents the QGW prior and the associated mixture model for Bayesian clustering, and Section 3.2 develops the Bayesian hierarchical model for treatment-contrast estimation, including an RMST-based causal estimand and an extension to single-arm trials with external control data. Section 4 evaluates the proposed methods under a range of simulation settings. Section 5 applies the framework to a prostate cancer case study combining data from a randomized trial, single-arm trials, and an observational study. Finally, Section 6 summarizes the main findings and discusses limitations and possible extensions.

2 Background: Geometric Weights Prior

Let $\mathcal{P}_{\mathbb{X}}$ be the set of probability measures on space $\mathbb{X}$ and Borel σ-field $\mathcal{X}$. The geometric weights (GW) prior is a random probability measure $\mathrm{P} \in \mathcal{P}_{\mathbb{X}}$ that can be defined in a constructive form (see Damien et al., 2013, Definition 14.1) as

$$\mathrm{P}(B) = \sum_{l=1}^{\infty} \omega_l \delta_{V_l}(B), \;\; \omega_l = \zeta(1-\zeta)^{l-1}, \;\; \zeta \sim \mathrm{Be}(a,b), \;\; a,b>0, \;\; B \in \mathcal{X}, \tag{1}$$

where the weights ω_l and locations V_l are random variables such that $\sum_{l=1}^{\infty} \omega_l = 1$ almost surely and $V_l \overset{ind.}{\sim} \mathrm{P}_0$ with $\mathrm{E}(\mathrm{P}) = \mathrm{P}_0$, a centering measure. Here, $\delta_{V_l}(B)$ denotes the Dirac measure centered at V_l. The mixture weights are defined through a stick-breaking mechanism, originally introduced for the Dirichlet process (DP) prior in Sethuraman (1994). Note that the DP prior arises from the stick-breaking construction with weights $\omega_1 = \zeta_1$ and $\omega_l = \zeta_l \prod_{r=1}^{l-1}(1-\zeta_r), \;\; l \geq 2$, where $\zeta_l \overset{i.i.d.}{\sim} \mathrm{Be}(1,\alpha)$, $l = 1, 2, \ldots$, a beta distribution with mean $1/(1+\alpha)$ and variance $\alpha/[(1+\alpha)^2(2+\alpha)]$.

While the DP prior involves an infinite number of random variables in its weights, the GW prior's weights depend on only a single beta random variable, meaning that in the context of stick-breaking construction, the stick is broken repeatedly using the same proportion, determined by that beta variable. Beyond its parsimonious structure, the GW prior also addresses a non-identifiability issue that can arise under the DP prior, in which the stick-breaking weights are not deterministically

ordered, unlike those in the GW prior. This issue renders DP-based models less appealing for clustering purposes, as it leads to *label switching* in posterior inference: the labels assigned to clusters are not identifiable. In response, post-hoc relabeling algorithms have been actively explored (Jasra et al., 2005; Sperrin et al., 2010; Papastamoulis, 2016). These approaches operate on MCMC output by producing permutations that rearrange component-specific parameter samples at each iteration, thereby facilitating comparability of parameters across posterior samples for each component. However, unlike finite mixture models, DP mixture models inherently assume an infinite number of components. Consequently, unless a truncation approximation of the DP is applied (Ishwaran and Zarepour, 2000; Ishwaran and James, 2001), implementing relabeling methods requires the additional step of selecting representative components from the MCMC output to fix the number of components. Even for finite mixture models or truncated DP mixture models with a fixed number of components, relabeling algorithms themselves are computationally intensive and can become impractical in large-scale settings, such as models with many components or high-dimensional data. In contrast, the GW prior breaks the remaining portion of the stick by a fixed fraction at each step, yielding naturally ordered weights from largest to smallest and thereby reducing the label-switching issue arising from the exchangeability and non-identifiability of mixture components.

Unlike the DP prior, the use of a single parameter to define the mixture weights in the GW prior may limit the flexibility in weight representation, as each weight is defined as a fixed fraction of the remaining stick. Consequently, a single underlying cluster may be erroneously split into multiple fragments, or distinct clusters may be incorrectly merged, distorting both the number of clusters and their sizes. In the context of integrative analysis of RWD and RCTs, this can cause RWD patients that should cluster with RCT patients to be assigned to separate groups and removed, leading to unintended loss of valuable data. Conversely, RWD patients that should not be grouped with RCT patients may be incorrectly clustered together with them. In the following section, we introduce a novel prior that preserves the ordered structure of the weights while enhancing flexibility in their configuration through additional parameters.

3 Methodology

3.1 Bayesian Model-based Clustering

In Section 3.1.1, we formulate a novel prior, termed the quasi-geometric weights (QGW) prior, and Section 3.1.2 constructs a mixture model based on the QGW prior.

3.1.1 Quasi-Geometric Weights Prior

In the stick-breaking construction of the GW prior, a constant fraction of the remaining stick is assigned as a weight. The fixed fraction in the GW prior can be overly restrictive when the true weights we aim to represent do not follow the terms of a geometric series. For example, consider a case with two underlying clusters of the same size (equal weight). A prior fraction of 0.5 can adequately represent one of the target weights, but the remaining 0.5 weight would be recovered by several smaller components, such as 0.25, 0.125, 0.0625, and so on. As we will show later, the weights serve as the prior probabilities that an observational unit is assigned to each component. Consequently, because the fixed fraction induces more small-weight components than it should, it may lead to overclustering – that is, the formation of more clusters (mixture components containing at least one unit) than truly exist. Our quasi-geometric weights prior generalizes the GW prior by

incorporating additional parameters that relax the assumption of a fixed stick-breaking fraction. Denoting some initial fractions by v_l for $l = 1, \ldots, M$, the weights ω_l are defined as:

$$\omega_l = \begin{cases} v_l \sim \text{Be}(a, b) \text{ for } l = 1 \\ v_l \prod_{r=1}^{l-1}(1 - v_r), \;\; v_l | v_{l-1} \sim \text{Be}(a, b) \text{ with } v_l \in \left(0, \min\left(1, \frac{v_{l-1}}{1-v_{l-1}}\right)\right) \text{ for } l = 2, 3, \ldots, M \\ v_M \left[\prod_{r=1}^{M-1}(1 - v_r)\right](1 - v_M)^{l-M} \text{ for } l = M+1, M+2, \ldots. \end{cases} \tag{2}$$

The formal definition of the QGW prior follows the same structure as the GW prior in (1), with the fixed stick-breaking fraction ζ replaced by v_l for the first M components in (2). The location parameters V_l are still sampled from P_0. The weights remain valid, with $\omega_l \in (0, 1)$ for $l = 1, \ldots$ and $\sum_{l=1}^{\infty} \omega_l = 1$; see Proposition S1 in Supplementary Material E.

This weight formulation demonstrates that the QGW prior retains the stick-breaking mechanism for weight generation, augmented by $M - 1$ additional parameters to enhance adaptability in the early break lengths. Regarding the upper bounds of a fraction, letting $v_l \to 1$ assigns essentially all of the remaining stick, $\prod_{r=1}^{l-1}(1 - v_r)$, to component l. Thus, the prior probabilities for cluster allocation become concentrated on the first (largest) l components. By allowing one component to absorb most of the remaining weight, the QGW prior permits subsequent tail weights to become negligible, thereby reducing the potential for overclustering. Since $v_{l-1} \geq 0.5$ is required for $\min\left(1, \frac{v_{l-1}}{1-v_{l-1}}\right) = 1$, the remaining weight satisfies $\prod_{r=1}^{l-1}(1 - v_r) \leq v_{l-1} \prod_{r=1}^{l-2}(1 - v_r) = \omega_{l-1}$. Therefore, the inequality $\omega_l < \omega_{l-1}$ holds, which is necessary for maintaining the weight ordering. On the other hand, as $v_l \to [v_{l-1}/(1 - v_{l-1})]$, the corresponding weight $\omega_l = v_l \prod_{r=1}^{l-1}(1 - v_r)$ increases toward $v_{l-1} \prod_{r=1}^{l-2}(1 - v_r) = \omega_{l-1}$ while staying strictly less than ω_{l-1}.

For $l = 2, \ldots, M$, the fractions v_l follow a truncated beta distribution, denoted by $\text{TBe}(a, b; \mathcal{L}, \mathcal{U})$, where $v_l \in (\mathcal{L}, \mathcal{U})$, $\mathcal{L} = 0$ and $\mathcal{U} = \min\left(1, \frac{v_{l-1}}{1-v_{l-1}}\right)$. When $\mathcal{U} = 1$, this reduces to the standard beta distribution, i.e., $\text{TBe}(a, b; 0, 1) = \text{Be}(a, b)$. The upper bound $\mathcal{U}$ is determined by v_{l-1}: if $v_{l-1} \geq 0.5$, then $\mathcal{U} = 1$; otherwise, $\mathcal{U} = v_{l-1}/(1 - v_{l-1})$. Although the beta distribution provides a more general specification, we restrict our attention in this paper to the special case $\text{Be}(1, 1) = \text{Unif}(0, 1)$, and hence $\text{TBe}(1, 1; \mathcal{L}, \mathcal{U}) = \text{Unif}(\mathcal{L}, \mathcal{U})$, for the following reasons. First, the resulting uniform prior assigns equal density over the admissible interval $(\mathcal{L}, \mathcal{U})$ for v_l, rather than favoring any subregion of this interval a priori. More importantly, this prior specification is key to tractable inference for v_l. In contrast, with a truncated beta prior having general hyperparameters, v_l determines the upper bound of the prior support for v_{l+1}, introducing a normalizing constant involving the regularized incomplete beta function into the full conditional distribution of v_l. As a result, the closed-form update used under the uniform specification is no longer available. Details of posterior inference under the model introduced in the following section, using this uniform prior specification, are provided in Supplementary Material A.1.2.

The number of fractions M is user-specified. In the absence of prior information, we use $M = \lfloor \sqrt{n/2} \rfloor$ as a practical choice motivated by a commonly used sample-size-based heuristic in clustering. As a diagnostic, one of the fractions v_l should be sufficiently close to 1 to absorb most of the remaining mass, yielding substantially smaller subsequent weights. If this does not occur for a given M, we recommend refitting the model with a larger value. We assess sensitivity to this choice in the simulation study, finding that results remain stable provided M is chosen sufficiently large; additionally, halving and doubling M in the case study produces identical cluster assignments.

Lastly, the QGW prior can be viewed as a flexible generalization of the GW prior. The GW

prior assumes a single stick-breaking fraction, which induces geometrically decaying weights at a fixed rate. The QGW prior instead introduces a finite set of fractions $\{v_l : l = 1, \ldots, M\}$, allowing the decay pattern of the weights to adapt more flexibly. When $M = 1$, only one fraction is used across all components, and the resulting weights reduce to those of the GW prior. Hence, the GW prior is nested within the QGW prior as a special case. It is also related to the DP stick-breaking construction, which is recovered when the ordering restriction is removed and infinitely many fractions are independently distributed as $\text{Be}(1, b)$ (i.e., setting $a = 1$ and $\mathcal{U} = 1$). As a middle ground between the GW and DP priors, the QGW prior inherits desirable properties from both: it retains the ordered-weight structure that helps mitigate label-switching issues while supporting richer weight configurations.

3.1.2 Mixture Model with QGW Prior

Mixture models are a well-established probabilistic, model-based approach to clustering and arguably the most widely used framework in Bayesian clustering (Fruhwirth-Schnatter et al., 2019; Bouveyron et al., 2019; Fraley and Raftery, 2002; Wade, 2023). Data are assumed to be drawn from a convex combination of components, which are, in general, densities from a well-known parametric distribution family, where each mixture component corresponds to a cluster. Let $\text{QGW}(G_0, a, b)$ denote the quasi-geometric weights prior with centering measure G_0 and stick-breaking parameters a and b governing the sampling fractions. With the constructive form of the QGW prior, $\text{P}(B) = \sum_{l=1}^{\infty} \omega_l \delta_{V_l}(B)$, our Bayesian nonparametric model is defined as

$$\begin{aligned} p(\boldsymbol{z}|G) &= \int k(\boldsymbol{z}|\boldsymbol{\theta}) dG(\boldsymbol{\theta}), \;\; G \sim \text{QGW}(G_0, a, b) \\ &= \sum_{l=1}^{\infty} \omega_l k(\boldsymbol{z}|V_l), \;\; V_l \sim G_0, \;\; l = 1, 2, \cdots, \end{aligned} \tag{3}$$

where the basis function $k(\boldsymbol{z}|\cdot)$ is the probability density or mass function of a parametric distribution for univariate $\boldsymbol{z}$, or the joint probability density and/or mass function for multivariate $\boldsymbol{z}$. Here, $\boldsymbol{\theta}$ is a generic element of the kernel parameter space, and V_l is the component-specific parameter (atom) for mixture component l. The weights ω_l are formulated as in (2); specifically, under our prior specification $a = b = 1$, they are given by:

$$\omega_l = \begin{cases} v_l \sim \text{Unif}(0, 1) \text{ for } l = 1 \\ v_l \prod_{r=1}^{l-1}(1 - v_r), \;\; v_l | v_{l-1} \sim \text{Unif}\Big(0, \min\big(1, \frac{v_{l-1}}{1 - v_{l-1}}\big)\Big) \text{ for } l = 2, 3, \ldots, M \\ v_M \Big[\prod_{r=1}^{M-1}(1 - v_r)\Big](1 - v_M)^{l-M} \text{ for } l = M+1, M+2, \ldots. \end{cases} \tag{4}$$

For the basis function $k(\boldsymbol{z}|\cdot)$, we employ normal, Poisson, and Bernoulli distributions for common data types: continuous, count, and categorical variables, respectively. These distribution choices facilitate posterior inference by allowing G_0 to be specified as conjugate priors for the corresponding distribution parameters.

Consequently, the basis function is defined as $k(\boldsymbol{z}|V_l) = \text{N}(\boldsymbol{z_1}|\mu_l, \sigma_l^2)\text{Pois}(\boldsymbol{z_2}|\lambda_l)\text{Bern}(\boldsymbol{z_3}|\pi_l)$ for $l = 1, 2, \ldots$, where $V_l = (\mu_l, \sigma_l^2, \lambda_l, \pi_l)$. Here, $\text{N}(\boldsymbol{z_1}|\mu, \sigma^2)$ denotes the probability density function of the normal distribution with mean μ and variance σ^2, while $\text{Pois}(\boldsymbol{z_2}|\lambda)$ and $\text{Bern}(\boldsymbol{z_3}|\pi)$ denote the probability mass functions of the Poisson and Bernoulli distributions with means λ and π, respectively. If $\boldsymbol{z_1} = (z_{11}, \ldots, z_{1p})$ (and likewise $\boldsymbol{z_2}$ and $\boldsymbol{z_3}$) is a vector representing multiple

(p) variables of a given data type, we assume that the elements are independent and identically distributed. Accordingly, the associated density (or mass) functions are given by

$$k(\boldsymbol{z}|V_l) = \prod_{q=1}^{p} \mathrm{N}(z_{1q}|\mu_{lq}, \sigma_{lq}^2) \prod_{q=1}^{p'} \mathrm{Pois}(z_{2q}|\lambda_{lq}) \prod_{q=1}^{p''} \mathrm{Bern}(z_{3q}|\pi_{lq}),$$

for vectors $\boldsymbol{z_1}$, $\boldsymbol{z_2}$, and $\boldsymbol{z_3}$ of lengths p, p', and p'', respectively.

We provide details of posterior inference, including the hierarchical representation of the model and MCMC implementation, in Supplementary Material A.1, where we assume that the data comprise all three types of variables to illustrate the corresponding posterior updates.

We employ this model as the first stage of integrative data analysis combining RCTs and RWD. As a nonparametric approach, it helps mitigate bias arising from potential model misspecification. Additionally, being an infinite mixture model, it offers two key advantages. First, it avoids the need to prespecify the number of mixture components. In finite-mixture approaches, sensitivity to this choice may arise, motivating alternative methods that randomize the number of components but typically require substantially more complex inference procedures. Second, it alleviates the issue of label switching, which is unavoidable in finite mixtures due to the unordered nature of component weights. For instance, the final component is often implicitly defined as the remainder (i.e., one minus the sum of the preceding weights), complicating label identification. While non-identifiability can still occur in infinite mixture models, we address this challenge through the use of the QGW prior described earlier.

The goal of this stage is to carefully select external data that resemble the clinical trial data; therefore, clusters consisting solely of RWD observations are excluded from the second-stage analysis. Because such clusters lack representation from the target RCT population, they cannot cannot directly inform inference on treatment contrasts for the RCT-supported population. Within each remaining cluster, RCT and RWD observations are expected to show more similar covariate distributions, which helps mitigate covariate shift and promotes baseline comparability. However, even among RCT data, patient characteristics may differ substantially and be clustered by unexpected factors – for example, variation across medical institutions in a multicenter research trial. By placing this clustering step at the beginning of the analysis pipeline, we establish a foundation for downstream analyses that can account for heterogeneity across latent clusters. This motivates the hierarchical model introduced in the following section, which captures not only differences between RCT and RWD sources, but also variation across the estimated clusters.

3.2 Bayesian Hierarchical Model

We herein introduce a hierarchical model developed for time-to-event data collected from multiple data sources. This second-stage model takes the Stage 1 clustering results as input. The clustering step is intended to identify subgroups of patients with similar baseline prognostic profiles across data sources. Borrowing is then restricted to patients within the same cluster, thereby reducing the risk of inappropriate information sharing across clinically heterogeneous populations. The clustering procedure does not guarantee comparability between RCT and RWD individuals; rather, it aims to partition patients into more homogeneous subgroups with respect to measured prognostic characteristics. Subsequent borrowing is therefore performed within clusters under the assumption that residual between-source heterogeneity is adequately captured by the Stage 2 model specified

below.

Let $i_{s,d}$ be the i-th observational unit, $i = 1, \ldots, n_{s,d}$, in cluster $s \in \mathcal{S}_d$ from dataset d, where $n_{s,d}$ is the number of units in the corresponding cluster-dataset pair; $\mathcal{S}_d = \{s^* \in \{1, \ldots, S\} : n_{s^*,d} > 0\}$ is the set of clusters represented in dataset d, and S is the total number of retained clusters from Stage 1 that contain at least one RCT observation. We allow the set of clusters $\mathcal{S}_d$ to vary across datasets, because a cluster may contain units from only a subset of datasets. For the model description below, we assume two RCT datasets (studies) and one real-world dataset, such that $d \in \{\text{RCT1}, \text{RCT2}, \text{RWD}\}$. However, the model is applicable to any number of RCT and RWD studies. Let $(t_{i_{s,d}}, \delta_{i_{s,d}})$ denote the observed survival data, where $t_{i_{s,d}}$ is the event or censoring time and $\delta_{i_{s,d}}$ is the censoring indicator (0 for censoring, 1 for an event). Let $x_{i_{s,d}}$ denote the treatment indicator, equal to 1 for the experimental arm (EA) and 0 for standard of care (SoC). Participants in the RCT datasets receive either EA or SoC, whereas those in the RWD dataset receive SoC only.

Our primary objective is not to transport RCT treatment-arm contrasts to the entire real-world population but rather to improve estimation of these contrasts within the population represented by the RCT. This objective reflects the fact that RWD may differ systematically from RCT populations due to differences in patient selection, outcome ascertainment, and follow-up. Consequently, the proposed framework uses RWD only as a supplemental source of control-arm information while preserving inference within the RCT-supported population. To this end, RWD are incorporated only within clusters containing both RCT and RWD observations. Clusters consisting exclusively of RWD observations are excluded from subsequent analyses. The target population of the proposed framework is therefore the union of individuals represented by the retained clusters. Consequently, inference is restricted to the RCT-supported population and should not be interpreted as representing treatment-arm contrasts in the overall combined RCT and RWD population.

The proposed Bayesian hierarchical model is built on the Cox proportional hazards framework, in particular the exponential proportional hazards model, and is defined as follows:

$$
\begin{aligned}
p\Big(\{(t_{i_{s,d}}, \delta_{i_{s,d}})\}|\Theta\Big) &\propto \prod_{d\in\{\text{RCT1,RCT2,RWD}\}} \prod_{s\in\mathcal{S}_d} \prod_{i=1}^{n_{s,d}} h(i,s,d)^{\delta_{i_{s,d}}} \exp\Big(-t_{i_{s,d}} h(i,s,d)\Big), \\
h(i,s,d) \equiv h(x_{i_{s,d}};\Theta) &= \begin{cases} \exp\left(\zeta_s + \zeta_{s,\text{RCT}} + x_{i_{s,d}}\gamma_{s,d}\right) \text{ if } d = \text{RCT1} \\ \exp\left(\zeta_s - \zeta_{s,\text{RCT}} + x_{i_{s,d}}\gamma_{s,d}\right) \text{ if } d = \text{RCT2} \\ \exp\left(\zeta_s + \zeta_{s,d}\right) \text{ if } d = \text{RWD} \end{cases} \qquad (5) \\
\zeta_{s,\text{RCT}}|\psi^2 &\overset{i.i.d.}{\sim} \text{N}(0,\psi^2) \text{ for } s = 1,\ldots,S \\
\zeta_{s,d}|\chi^2 &\overset{i.i.d.}{\sim} \text{N}(0,\chi^2) \text{ for } s = 1,\ldots,S,\ \ d \in \{\text{RWD}\} \text{ such that } n_{s,d} > 0 \\
\zeta_s|\phi^2 &\overset{i.i.d.}{\sim} \text{N}(0,\phi^2) \text{ for } s = 1,\ldots,S \\
\gamma_{s,d}|\gamma_s,\upsilon^2 &\overset{ind.}{\sim} \text{N}(\gamma_s,\upsilon^2) \text{ for } s = 1,\ldots,S,\ \ d \in \{\text{RCT1},\text{RCT2}\} \text{ such that } n_{s,d} > 0 \\
\gamma_s|\gamma,\tau^2 &\overset{i.i.d.}{\sim} \text{N}(\gamma,\tau^2) \text{ for } s = 1,\ldots,S \\
\gamma|\sigma^2 &\sim \text{N}(0,\sigma^2) \\
(\sigma^2,\tau^2,\upsilon^2,\phi^2,\chi^2,\psi^2) &\overset{i.i.d.}{\sim} \text{IG}(a,b),
\end{aligned}
$$

where $\Theta = \Big(\{\zeta_{s,\text{RCT}}\}, \{\zeta_{s,\text{RWD}}\}, \{\zeta_s\}, \{\gamma_{s,\text{RCT1}}\}, \{\gamma_{s,\text{RCT2}}\}\Big)$.

The parameters $\gamma_{s,d}$, γ_s, and γ have distinct roles in the hierarchical model. The parameter $\gamma_{s,d}$

denotes the cluster- and dataset-specific log hazard ratio (log-HR) associated with the experimental arm in RCT dataset d and cluster s. These lower-level parameters allow the treatment-arm log-HR contrast to vary across both clusters and RCT studies. The higher-level parameters γ_s and γ are model-based hierarchical summaries of the lower-level log-HR contrasts. Specifically, γ_s summarizes the cluster-specific log-HR across RCT datasets, whereas γ represents the grand mean log-HR across the retained RCT-supported clusters. Thus, $\gamma_{s,d}$ captures local cluster- and dataset-level heterogeneity, while γ_s and γ provide cluster-level and overall log-HR summaries, respectively. Supplementary Material F presents asymptotic considerations for the estimation of the top-level parameter γ.

Our key quantities of interest are hazard ratios (HRs). Although HRs are non-collapsible and may be difficult to interpret causally (Hernán, 2010), they remain commonly reported summary measures of treatment benefit in oncology trials, including in settings with time-varying treatment effects or covariate-dependent hazards. Since HR-based summaries can be challenging to interpret clinically, we also consider restricted mean survival time (RMST)-based contrasts, which have been suggested as suitable causal estimands in survival analysis (Uno et al., 2014; Ni et al., 2021) and can be interpreted as the average survival time gained or lost up to a prespecified horizon. We present the RMST under the proposed model using the *potential outcomes framework* in Section 3.2.1.

The baseline hazard parameters – $\zeta_s, \zeta_{s,\text{RCT}}$, and $\zeta_{s,\text{RWD}}$ – are also allowed to vary across clusters. ζ_s represents the common RCT effect; it is defined as the average of RCT1's $(\zeta_s + \zeta_{s,\text{RCT}})$ and RCT2's $(\zeta_s - \zeta_{s,\text{RCT}})$. Its exponential corresponds to the geometric mean of the baseline hazards for RCT1 and RCT2, such that $\exp(\zeta_s) = \sqrt{\exp\left(\zeta_s + \zeta_{s,\text{RCT}}\right)\exp\left(\zeta_s - \zeta_{s,\text{RCT}}\right)}$. The parameter $\zeta_{s,\text{RWD}}$ for RWD is introduced to capture discrepancies from the RCTs' averaged baseline hazard. Such discrepancies may arise due to potential disparities between the RCT control groups and the external control (RWD), which could result from outcome incomparability.

The sum $\zeta_s + \zeta_{s,d}$ can be reparameterized as $\zeta'_{s,d} = \zeta_s + \zeta_{s,d}$, with $\zeta'_{s,d}|\zeta_s, \chi^2 \overset{ind.}{\sim} \text{N}(\zeta_s, \chi^2)$. This representation helps clarify the role of the variance parameter χ^2 in the estimation of ζ_s, with χ^2 controlling the extent to which the RWD contributes to this estimation. Smaller values of χ^2 induce stronger borrowing from the RWD, whereas larger values weaken this contribution. When $\chi^2 = 0$, we have $\zeta'_{s,d} = \zeta_s$, so replacing $\zeta'_{s,d}$ with ζ_s implies that the RWD likelihood directly informs the estimation of ζ_s. For nonzero χ^2, the RWD likelihood informs ζ_s only indirectly through the conditional prior linking $\zeta'_{s,d}$ to ζ_s. As χ^2 increases, the ζ_s-dependent kernel of this prior, $\exp\left(-(\zeta'_{s,d} - \zeta_s)^2/(2\chi^2)\right)$, becomes increasingly flat as a function of ζ_s, weakening the connection between $\zeta'_{s,d}$ and ζ_s. Consequently, posterior inference for ζ_s becomes increasingly less dependent on $\zeta'_{s,d}$ and, in turn, on the contribution of the RWD likelihood. A more formal characterization of the role of χ^2, based on a local Gaussian approximation to the RWD likelihood for $\zeta'_{s,d}$, is provided in Supplementary Material G. The derivation shows that larger values of χ^2 progressively reduce the contribution of the RWD likelihood to posterior inference for ζ_s, with this contribution vanishing in the limit as $\chi^2 \to \infty$. Because χ^2 is estimated from the data rather than user-specified, the degree of borrowing adapts to the compatibility between the RWD and RCT baseline hazards as informed by the data.

The proposed modeling framework can also be extended to single-arm trials (SATs), in which concurrent control data are unavailable and control information may instead be obtained from external sources, such as historical trial data or RWD. Section 3.2.2 presents the corresponding formulation for SATs with external control data.

Although randomization makes it unlikely that a cluster-specific subgroup within an RCT will

contain patients from only one treatment arm, this may occur by chance in small clusters. If all patients in a cluster-RCT subgroup receive SoC, the data provide no direct information about the cluster-specific treatment effect, and inference is therefore informed by the hierarchical prior through borrowing across other clusters and/or RCTs. Conversely, when such a subgroup contains only EA patients, the baseline hazard components ζ_s and $\zeta_{s,\text{RCT}}$ cannot be separately identified from the log-HR parameter $\gamma_{s,d}$ within that subgroup. In this setting, a structural modification is needed for identifiability. Specifically, we impose a common baseline hazard component across the two RCTs, denoted by $\tilde{\zeta}_s$, provided that the baseline hazard component is identifiable in the other RCT. If neither RCT provides information on the baseline component for cluster s (i.e., all patients in both RCTs receive EA), the model reduces to the single-arm specification described in Section 3.2.2 for that cluster.

Lastly, the model can be generalized to multiple studies, including N_T RCTs and N_W RWD sources, where the set of labels that d can take is expanded to $d \in \{\text{RCT1}, \ldots, \text{RCT}N_T\} \cup \{\text{RWD1}, \ldots, \text{RWD}N_W\}$. The RCT baseline hazard deviation, originally denoted $\zeta_{s,\text{RCT}}$, is modified to $\zeta_{s,r,\text{RCT}}$ for $r = 1, , 2, \ldots, (N_T - 1)$ in a manner that ensures $\exp(\zeta_s)$ represents the average baseline hazard across RCT studies. For instance, one may assign $\zeta_s + \zeta_{s,r,\text{RCT}}$ to the first $N_T - 1$ RCT studies, and $\zeta_s - \sum_{r=1}^{N_T-1} \zeta_{s,r,\text{RCT}}$ to the last RCT. This construction establishes identifiability and preserves the interpretation of $\exp(\zeta_s)$ as the representative baseline hazard across RCT studies. The indexing of d in $\zeta_{s,d}$ and $\gamma_{s,d}$ is updated to reflect dataset-specific effects: $d \in \{\text{RWD1}, \ldots, \text{RWD}N_W\}$ in $\zeta_{s,d}$ and $d \in \{\text{RCT1}, \ldots, \text{RCT}N_T\}$ in $\gamma_{s,d}$. For greater flexibility across datasets of the same type, the variance hyperparameters of the priors on $\zeta_{s,r,\text{RCT}}$, as well as those on $\zeta_{s,d}$, can be extended to ψ_r^2 and χ_d^2, respectively.

In Supplementary Material A.2, we provide the posterior inference method for this hierarchical model, in which all full conditional distributions are available in closed form under normal and inverse-gamma priors. The hyperparameters for the inverse-gamma priors are set to $a = 2$ and $b = 10$ as in the QGW mixture model.

3.2.1 Restricted Mean Survival Time as a Causal Estimand

Let $X_{i_{s,d}} \in \{0, 1\}$ denote the treatment arm for the i-th individual in cluster s of RCT dataset d, where $X_{i_{s,d}} = 1$ indicates the experimental arm and $X_{i_{s,d}} = 0$ indicates standard of care. For $x = 0, 1$, let $T_{i_{s,d}}(x)$ denote the potential event time that would occur under treatment arm x.

Following Lu and Ni (2025), we define the RMST for treatment arm x as the area under the survival function up to a prespecified time horizon τ, chosen based on clinical knowledge. For a fixed time point t, let $S_{s,d}^{(x)}(t) = \Pr(T_{i_{s,d}}(x) > t)$ denote the survival probability associated with treatment arm x. Under the proposed model in (5), for $s \in \mathcal{S}_d$ and $d \in \{\text{RCT1}, \text{RCT2}\}$, this survival probability is given by $S_{s,d}^{(x)}(t) = \exp\big(-h(x; \Theta)t\big)$, where the hazard function $h(x; \Theta)$ is specified differently for each RCT dataset d. The RMST for treatment arm x is obtained by integrating the survival function from 0 to τ, yielding

$$\text{RMST}_{s,d}^{(x)}(\tau) = \int_0^\tau \exp\big(-h(x; \Theta)t\big)dt = \frac{1 - \exp(-h(x; \Theta)\tau)}{h(x; \Theta)}.$$

The corresponding RMST-based treatment effect can then be defined as $\Delta_{\text{RMST}_{s,d}}(\tau) = \text{RMST}_{s,d}^{(1)}(\tau) - \text{RMST}_{s,d}^{(0)}(\tau)$. Under this estimation framework, the posterior mean of the treatment effect is obtained by averaging $\Delta_{\text{RMST}_{s,d}}(\tau)$ over the MCMC samples, where $\Delta_{\text{RMST}_{s,d}}(\tau)$ is

computed at each iteration by substituting the posterior draw of the model parameters into the expression for $\text{RMST}_{s,d}^{(x)}(\tau)$.

For the causal interpretation, we invoke identifying assumptions. These assumptions are standard in potential-outcomes analyses but are generally not fully verifiable from observed data. Consequently, causal interpretation depends on the plausibility of the design and subject-matter assumptions underlying the analysis. First, consistency is assumed, so that the observed event time equals the potential event time under the treatment actually received. Second, the stable unit treatment value assumption (SUTVA) is assumed within each RCT study, implying that one participant's outcome is unaffected by the treatment assignments of other participants (no interference) and that each treatment is well-defined with no hidden variations within treatment arms. Third, treatment positivity is assumed within each retained cluster and RCT dataset. Fourth, censoring is assumed to be independent of the potential event times conditional on data source and cluster. In addition, because RWD individuals contribute external control information, we assume that, within retained clusters and after accounting for the RWD-specific baseline hazard deviation parameters, RWD control outcomes are sufficiently comparable to RCT control outcomes to support borrowing. This comparability assumption cannot be verified empirically using observed data alone, because residual differences between RCT and RWD control outcomes may remain due to unmeasured prognostic factors. Therefore, it requires strong clinical and methodological justification, including comparability of patient eligibility criteria, outcome definitions, follow-up procedures, treatment settings, and clinical practice patterns across data sources.

3.2.2 Extension to Single-Arm Clinical Trial

The primary difference from the proposed RCT model lies in the hazard function, particularly in the baseline hazard: the parameter $\zeta_{s,\text{RCT}}$, which accounted for baseline hazard deviations across RCT datasets in (5), is omitted, leaving only ζ_s. Because the SAT baseline hazard information is borrowed entirely from the external RWD controls, estimating the cross-trial baseline discrepancy parameter for SATs is neither meaningful nor separately identifiable. Formally, the hazard function for the SAT extension is defined as follows:

$$h(i,s,d) \equiv h(x_{i_{s,d}};\Theta) = \begin{cases} \exp\left(\zeta_s + x_{i_{s,d}}\gamma_{s,d}\right) \text{ if } d = \text{SAT1}, \text{SAT2} \\ \exp\left(\zeta_s\right) \text{ if } d = \text{RWD}, \end{cases} \tag{6}$$

where $\exp(\zeta_s)$ represents the baseline hazard for the RWD, which also serves as the baseline hazard for the SATs. Crucially, in the absence of RWD, ζ_s and $\gamma_{s,d}$ are not separately identifiable. This is because $x_{i_{s,d}} = 1$ for all patients in the SAT groups, causing ζ_s and $\gamma_{s,d}$ to collapse into a single additive term. Consequently, many different combinations of ζ_s and $\gamma_{s,d}$ yield the same hazard $h(i,s,d)$ for the SATs, demonstrating that ζ_s can be identified separately from $\gamma_{s,d}$ only through the incorporation of RWD external controls. The prior distributions specified in (5) remain applicable to the model parameters in (6).

The model can be generalized to accommodate N_T SATs and N_W RWD sources. The corre-

sponding hazard function $h(i, s, d)$ is defined as

$$h(i,s,d) = \begin{cases} \exp\left(\zeta_s + x_{i_{s,d}}\gamma_{s,d}\right) \text{ if } d \in \{\text{SAT1}, \ldots, \text{SAT}N_T\} \\ \exp\left(\zeta_s + \zeta_{s,d}\right) \text{ if } d \in \{\text{RWD1}, \ldots, \text{RWD}N_W - 1\} \\ \exp\left(\zeta_s - \sum_{d \in \{\text{RWD1},\ldots,\text{RWD}N_W-1\}} \zeta_{s,d}\right) \text{ if } d = \text{RWD}N_W. \end{cases}$$

Here, $\exp(\zeta_s)$ represents the average baseline hazard across RWD sources, defined as the geometric mean of their respective baseline hazards, with $\zeta_{s,d}$ used for deviations from this average.

4 Simulation Study

4.1 Comparison between GW and QGW in Clustering

In this section, we compare the clustering performance of the QGW prior model with that of the GW prior model. Synthetic data are generated from three different normal distributions: $\text{N}(0, 1)$, $\text{N}(5, 1)$, and $\text{N}(10, 1)$, with corresponding sample sizes $M_1 = 70$, $M_2 = 80$, and $M_3 = 50$. Both the GW and QGW models follow the same mixture formulation as in (3), with the difference being the prior on the mixing distribution G: the GW prior and the QGW prior. In both models, the basis functions are specified as normal densities $k(\boldsymbol{z}|\mu_l, \sigma_l^2) = \text{N}(\boldsymbol{z}|\mu_l, \sigma_l^2)$.

We assess performance based on the estimated number and size of clusters, where the true data structure consists of 3 clusters with sizes $(70, 80, 50)$. The cluster assignment for each unit $i = 1, \ldots, 200$ is estimated by the posterior mode of the latent variable u_i in (S2) of Section A.1.1 of the Supplementary Material. First, the number of unique values among these 200 posterior modes serves as the estimated number of clusters. We repeat this process over 200 simulations, each involving data generation, model fitting, and cluster number estimation. The results, shown in the left panel of Figure 1, indicate that the GW model correctly estimates the number of clusters in only 135 out of 200 runs, whereas the QGW model achieves accurate estimates in 198 runs.

Next, we evaluate cluster sizes using the mean squared error (MSE), computed as

$$\Big(\sum_{b=1}^{200}(\hat{M}_1^{(b)} - M_1)^2/200, \sum_{b=1}^{200}(\hat{M}_2^{(b)} - M_2)^2/200, \sum_{b=1}^{200}(\hat{M}_3^{(b)} - M_3)^2/200\Big).$$

For a given simulation b, $\hat{M}_l^{(b)}$ denotes the estimated number of units assigned to cluster l, where cluster labels are determined by the associated location parameter estimate (posterior mean), $\hat{\mu}_l^{(b)}$. Specifically, the label l is assigned such that $\hat{\mu}_l^{(b)}$ is closer in Euclidean (L_2) distance to the true location parameter μ_l than to any other true parameter μ_s, where $s \neq l$ and $s \in \{1, 2, 3\}$. To ensure that exactly three cluster sizes are estimated for each replicate b (and can thus be paired with the corresponding true cluster sizes, M_s, $s = 1, 2, 3$), the MSE is computed exclusively using replicates in which both models correctly identify the true number of clusters. Conditioning the MSE calculation on successful recovery of the true number of clusters isolates the accuracy of cluster-size estimation from errors in estimating the number of occupied clusters. We continuously generate datasets until we accrue 200 such qualifying replicates. The right panel of Figure 1 clearly demonstrates the superior performance of the QGW model in estimating cluster sizes, as evidenced by its substantially lower MSE values.

To further assess agreement between the estimated and true cluster assignments, we computed

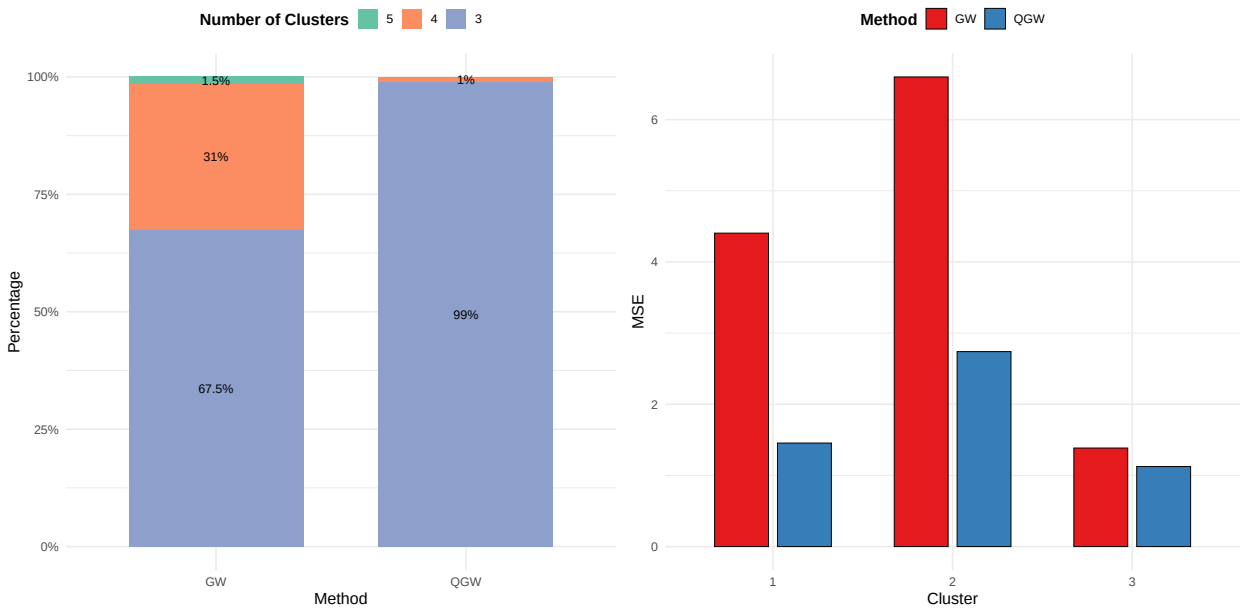


Figure 1: Stacked bar plot showing the percentage of estimated cluster numbers under each method (left), and bar plot of the mean squared error in cluster sizes for simulations that yield exactly three clusters (right). GW: the GW prior model; QGW: the QGW prior model.

the adjusted Rand index (ARI), which compares whether pairs of units are consistently grouped together or separated in the true and estimated partitions. Across the 200 simulations used to estimate the number of occupied clusters, the mean ARI was 0.966 for GW and 0.972 for QGW, indicating higher clustering agreement under QGW.

The results for the QGW model above are based on $M = 5$, considering that the underlying number of clusters is known to be 3, where the third stick-breaking fraction v_3 is expected to be close to 1, leading the third weight w_3 to capture nearly all of the remaining mass. Supplementary Material D presents the results of a sensitivity analysis on the choice of M, indicating that more conservative choices $M = 10, 15, 20$ do not compromise inferential accuracy.

As an additional simulation study, we consider more challenging scenarios: higher-dimensional covariates, weaker cluster separation, and correlated covariates. Across all scenarios, we assume four clusters with sizes $M_1 = 60$, $M_2 = 55$, $M_3 = 45$, and $M_4 = 40$. Let $s = 1, \ldots, 4$ index the true clusters. For each cluster s, data are generated from a multivariate normal distribution specified as $\mathrm{N}(\mu_s^{\mathrm{True}}, \Sigma_s^{\mathrm{True}})$. The three settings are defined as follows, where $\mathbb{I}_3$ denotes the 3×3 identity matrix. Across all scenarios, the cluster means are fixed at $\mu_1^{\mathrm{True}} = (-10, 4, 0)'$, $\mu_2^{\mathrm{True}} = (-5, 0, 6)'$, $\mu_3^{\mathrm{True}} = (0, -4, -3)'$, and $\mu_4^{\mathrm{True}} = (5, -8, 3)'$.

- Scenario 1: Independent covariates, with $\Sigma_s^{\mathrm{True}} = \mathbb{I}_3$ for all $s = 1, \ldots, 4$.
- Scenario 2: Independent covariates with weaker cluster separation, where $\Sigma_s^{\mathrm{True}} = 5\mathbb{I}_3$ for all $s = 1, \ldots, 4$. The larger within-cluster variance increases overlap among clusters.
- Scenario 3: Correlated covariates, with

Criterion	Scenario 1	Scenario 2	Scenario 3
Number of clusters correctly estimated (%)	93%	90.5%	87%
Cluster size estimation (MSE)	0/0/0/0	6/7/8/7	0/0/0/0
Adjusted Rand index	0.999	0.908	0.994

Table 1: Cluster recovery performance of the QGW model under three scenarios.

$$\Sigma_s^{\text{True}} = \begin{pmatrix} 1 & a_s & a_s b_s \\ a_s & 1 & b_s \\ a_s b_s & b_s & 1 \end{pmatrix}, (a_1, \ldots, a_4) = (0.2, 0.3, 0.4, 0.5) \text{ and } (b_1, \ldots, b_4) = (0.4, 0.5, 0.6, 0.7).$$

As in the previous example, we use 200 simulated datasets to estimate the number of clusters and compute the mean ARI. As shown in Table 1, increasing the number of covariates had little impact on clustering performance in the settings considered here, whereas reduced cluster separability leads to performance loss. In Scenario 2, where cluster separation is weakened by larger variances, performance decreases across all three measures relative to Scenario 1. Scenario 3 retains the variance settings of Scenario 2 but introduces correlation among the covariates, which also results in lower performance than in Scenario 1. These patterns become more pronounced when the variances are further increased in Scenario 2 and the correlation is strengthened in Scenario 3, as detailed in Table S1 in Supplementary Material H. Nevertheless, the QGW model continues to outperform the GW model, demonstrating a clear advantage in cluster recovery (Table S2 in Supplementary Material H).

4.2 Model Performance in Estimating the Overall Contrast γ

This section illustrates the proposed Bayesian hierarchical model defined in (5) using synthetic datasets and evaluates its performance in estimating the key parameter of interest, the overall treatment contrast γ, the mean squared error of the posterior mean across simulation replicates. A key advantage of combining RCT and RWD is the potential to improve estimation when the number of control-group patients in RCTs is limited, including cases in which no concurrent control patients are available, as in single-arm trials addressed by the model extension in (6). Accordingly, we present results for different proportions of control-group patients in RCTs.

To quantify the contribution of external data, we compare the proposed model with a counterpart that excludes the RWD component, such that $h(i, s, d)$ includes only RCT1 and RCT2, and $\zeta_{\text{s,RWD}}$ and χ^2 are removed from the model. Data are generated using the proportional hazards model with the underlying hazard functions:

$$h(i,s,d)^{\text{TRUE}} = \begin{cases} \exp\left(\zeta_s + \zeta_{s,\text{RCT}} + x_{i_{s,d}}\gamma_{s,d}\right) \text{ if } d = \text{RCT1} \\ \exp\left(\zeta_s - \zeta_{s,\text{RCT}} + x_{i_{s,d}}\gamma_{s,d}\right) \text{ if } d = \text{RCT2} \\ \exp\left(\zeta_s\right) \text{ if } d = \text{RWD}, \end{cases}$$

where $(\zeta_1, \zeta_2, \zeta_3, \zeta_4) = (-4, -5, -3, -2)$, $(\zeta_{1,\text{RCT}}, \zeta_{2,\text{RCT}}, \zeta_{3,\text{RCT}}, \zeta_{4,\text{RCT}}) = (2, 1, 3, 1)$, $\gamma_{s,d} \sim \text{N}(\gamma_s, 10)$ for $d \in \{\text{RCT1}, \text{RCT2}\}$, $\gamma_s \sim \text{N}(\gamma, 10)$ for s = 1,...,4, and $\gamma = 10$.

The grouping variable $x_{i_{s,d}}$ is sampled from Bern(p^*), where p^* denotes the probability of assignment to the treatment group (i.e., the group receiving the experimental treatment). To consider varying degrees of control-group scarcity in the RCTs, we use $p^* = 0.5$, 0.725, and 0.95,

Control-Group Scarcity (p^*)	0.5	0.725	0.95	0.5	0.725	0.95	0.5	0.725	0.95
Clustering Label	True	True	True	Estimated 1	Estimated 1	Estimated 1	Estimated 2	Estimated 2	Estimated 2
eRWD	12.20	12.44	26.35	12.48	13.02	26.77	15.99	17.44	33.29
iRWD	12.19	12.33	18.54	12.40	12.70	19.28	15.79	17.04	24.92

Table 2: MSE of the overall treatment contrast γ under different levels of control-group scarcity using true or estimated clustering labels. Estimated 1 and 2 correspond to different levels of cluster separation in the synthetic data, with $V^2 = 1$ and 5 for $\boldsymbol{z_{i_{s,d}1}}$, respectively. eRWD denotes the model excluding RWD, and iRWD the model including RWD.

representing balanced randomization, moderate control-group scarcity, and severe control-group scarcity with the smallest proportion of RCT control patients, respectively. Cluster sizes $n_{s,d}$ for $s = 1, \ldots, 4$ are set to $(n_{1,d}, n_{2,d}, n_{3,d}, n_{4,d}) = (10, 20, 10, 0)$ for $d =$ RCT1, $(10, 10, 15, 0)$ for $d =$ RCT2, and $(30, 15, 10, 20)$ for $d =$ RWD. Under these specifications, we generate 40 RCT1, 35 RCT2, and 75 RWD time-to-event observations $(t_{i_{s,d}}, \delta_{i_{s,d}})$, with randomly generated censoring times assigned to 20% of individuals within each cluster s and data source d.

To illustrate the entire procedure of our two-stage method using synthetic data, we begin with the clustering step, for which we also simulate covariates $(\boldsymbol{z_{i_{s,d}1}}, \boldsymbol{z_{i_{s,d}2}}, \boldsymbol{z_{i_{s,d}3}})$ from cluster-specific distributions as follows:

- For $s = 1$, $\boldsymbol{z_{i_{s,d}1}} \overset{i.i.d.}{\sim} \mathrm{N}(-10, V^2)$, $\boldsymbol{z_{i_{s,d}2}} \overset{i.i.d.}{\sim} \mathrm{Pois}(0.5)$, and $\boldsymbol{z_{i_{s,d}3}} \overset{i.i.d.}{\sim} \mathrm{Bern}(0.1)$;
- For $s = 2$, $\boldsymbol{z_{i_{s,d}1}} \overset{i.i.d.}{\sim} \mathrm{N}(-5, V^2)$, $\boldsymbol{z_{i_{s,d}2}} \overset{i.i.d.}{\sim} \mathrm{Pois}(1.5)$, and $\boldsymbol{z_{i_{s,d}3}} \overset{i.i.d.}{\sim} \mathrm{Bern}(0.3)$;
- For $s = 3$, $\boldsymbol{z_{i_{s,d}1}} \overset{i.i.d.}{\sim} \mathrm{N}(0, V^2)$, $\boldsymbol{z_{i_{s,d}2}} \overset{i.i.d.}{\sim} \mathrm{Pois}(2.5)$, and $\boldsymbol{z_{i_{s,d}3}} \overset{i.i.d.}{\sim} \mathrm{Bern}(0.6)$;
- For $s = 4$, $\boldsymbol{z_{i_{s,d}1}} \overset{i.i.d.}{\sim} \mathrm{N}(5, V^2)$, $\boldsymbol{z_{i_{s,d}2}} \overset{i.i.d.}{\sim} \mathrm{Pois}(3.5)$, and $\boldsymbol{z_{i_{s,d}3}} \overset{i.i.d.}{\sim} \mathrm{Bern}(0.9)$.

Values of $V^2 = 1$ and $V^2 = 5$ represent cases of well-separated clusters and weaker cluster separation, respectively. We adopt pooled notation that collapses the double-indexed structure $i_{s,d}$ into a single sequence i, such that $\boldsymbol{z_{i1}} = \boldsymbol{z_{i_{s,d}1}}$, $\boldsymbol{z_{i2}} = \boldsymbol{z_{i_{s,d}2}}$, $\boldsymbol{z_{i3}} = \boldsymbol{z_{i_{s,d}3}}$ for $i = 1, \ldots, n$, where $n = \sum_{s=1}^{4} \sum_{d \in \{\mathrm{RCT1,RCT2,RWD}\}} n_{s,d}$. These pooled covariates are used as input for the clustering step.

Table 2 presents the MSE of the overall treatment contrast over 500 replications under three levels of control-group scarcity, corresponding to three values of p^*. To empirically assess the impact of first-stage clustering on second-stage estimation, we considered both the true clustering labels and estimated labels obtained under two levels of cluster separation. The first set of estimated labels was obtained with $V^2 = 1$, yielding an ARI of 0.972, whereas the second was obtained under weaker separation with $V^2 = 5$, yielding an ARI of 0.708.

Across all levels of control-group scarcity and clustering accuracy considered, iRWD yields smaller MSEs for γ than eRWD, with the difference in MSE between the two models widening as the MSEs under both models increase. Furthermore, within both models, as the proportion of control-group patients decreases, the MSE of γ increases. This likely reflects less accurate estimation of ζ_s, $s = 1, 2, 3$, which is supported by Figure S2 of the Supplementary Material showing that reduced control-group size is associated with larger MSEs in the estimation of ζ_s. (This figure is restricted to the true cluster labels because the number of estimated clusters varies across the 500 replications, resulting in a varying number of estimated ζ_s parameters.) Furthermore, less accurate first-stage clustering adversely affects the estimation of γ: the MSEs under Estimated 2 ($V^2 = 5$)

Scenario	Constant shift			Unmeasured confounding		
Control-Group Scarcity (p^*)	0.5	0.725	0.95	0.5	0.725	0.95
eRWD	12.48	13.02	26.77	14.91	18.61	23.56
iRWD	12.40	12.77	25.90	15.11	19.24	24.50
iRWD with $\chi^2 = 0$	15.08	17.42	42.85	17.55	23.45	30.96

Table 3: MSE of the overall treatment contrast γ under two RCT-RWD discrepancy scenarios, three levels of control-group scarcity, and different specifications for incorporating RWD. Setting $\chi^2 = 0$ implies $\zeta_{s,\text{RWD}} = 0$. Results are based on the Estimated 1 clustering labels.

are larger than those under Estimated 1 ($V^2 = 1$), which are in turn larger than those obtained using the true labels.

The preceding study assumes comparability between the RCT and RWD control populations in terms of their baseline hazards, with the RWD baseline hazard $\exp(\zeta_s)$ matched to the geometric mean of the baseline hazards for RCT1 and RCT2. In practice, however, systematic differences may exist between RCT and RWD populations. We therefore consider examples that violate this assumption, under which incorporating RWD may introduce bias into estimation, thereby illustrating the utility of $\zeta_{s,\text{RWD}}$. We intentionally generate discrepancies between the RCT and RWD control populations through two mechanisms: a shift in the RWD log-baseline hazard and unmeasured confounding, representing two common threats to valid RCT-RWD integration – differences in baseline prognosis and residual confounding due to unmeasured variables, respectively. For the shift case, the RWD hazard $h(i, s, d)^{\text{TRUE}}$ is replaced by $\exp(\zeta_s + C_s)$, where C_s is drawn from a uniform distribution $\text{Unif}(-10, 10)$ in each of the 500 replications and remains constant across individuals within a replication. For the unmeasured confounding case, an additional variable $y_{i_{s,d}}$ is generated from different distributions for the RCT and RWD populations, with $y_{i_{s,\text{RCT1}}}, y_{i_{s,\text{RCT2}}} \overset{i.i.d.}{\sim} \text{N}(2, 1)$ and $y_{i_{s,\text{RWD}}} \overset{i.i.d.}{\sim} \text{N}(3, 1)$. The true hazard is then modified as $h(i, s, d)^{\text{TRUE}} \times \exp(y_{i_{s,d}}\beta)$ with $\beta = 3$. By treating $y_{i_{s,d}}$ as unobserved throughout both stages of the analysis, it acts as an unmeasured confounder that induces structural incomparability between the RCT and RWD control cohorts. Table 3 demonstrates that including $\zeta_{s,\text{RWD}}$ mitigates the loss of estimation accuracy that may occur when incorporating RWD in the presence of such discrepancies. Furthermore, when the discrepancy manifests as a constant shift, the model yields a lower MSE because $\zeta_{s,\text{RWD}}$ directly absorbs this systematic difference.

Overall, the numerical values were chosen to evaluate the operating characteristics of the proposed framework across a range of simulation settings, rather than to replicate a specific clinical application.

4.3 Model Comparison with the Power Prior Model

This section compares our approach with the propensity score-integrated power prior approach of Wang et al. (2019), a commonly used method that, similar to ours, stratifies patients. Specifically, this approach uses propensity scores to exclude incomparable RWD patients and forms strata based on quantiles of the propensity scores. Within each stratum, a power prior is used, with the power parameter determined by the degree of overlap between the propensity score distributions of the RCT and RWD units. Because this approach does not impose a hierarchical structure linking the stratum-specific parameters, we focus the comparison on the stratum-specific treatment contrasts

	Stronger separation ($V^2 = 1$)			Weaker separation ($V^2 = 5$)		
	$p^* = 0.5$	$p^* = 0.725$	$p^* = 0.95$	$p^* = 0.5$	$p^* = 0.725$	$p^* = 0.95$
BHM	3.89|4.27|6.68	4.01|5.16|7.19	6.30|9.18|12.70	4.42|7.94|13.72	4.43|9.70|16.75	7.59|15.87|26.31
PPM	9.07|19.48|26.34	9.47|22.33|32.98	10.73|24.83|36.53	8.89|20.41|31.56	9.39|23.07|38.57	10.58|25.87|44.29

Table 4: MSEs of γ_1, γ_2, and γ_3 under our Bayesian hierarchical model (BHM) and the power prior model (PPM) for two levels of cluster separation and three levels of control-group scarcity. Within each cell, the three values correspond to γ_1, γ_2, and γ_3, respectively.

γ_s, which are directly estimated under both approaches. Accordingly, we compare the MSEs of γ_s under the two methods.

For a common basis of comparison, we use the same simulated datasets as in the previous section and fix the number of clusters at the true value of three for both approaches. We therefore retain the 350 replications in which the first-stage procedure of our approach identifies exactly three clusters. Because the power prior approach does not estimate the number of clusters, three strata are specified using tertiles of the propensity scores for each of these replications. The power prior model is implemented using the `pwe.stratified.pp()` function in the `hdbayes` R package, with the two RCT datasets combined into a single RCT dataset. Table 4 shows that our Bayesian hierarchical model, which accommodates heterogeneity across datasets and clusters, yields smaller MSEs than the power prior model across all levels of control-group scarcity and cluster separation considered. Supplementary Table S3 reports the corresponding comparison under the two RCT-RWD discrepancy scenarios considered in Table 3 and further demonstrates the favorable performance of our model. Having demonstrated the benefits of the proposed approach across a range of simulated settings, we next evaluate its performance in a real-world application.

5 Case Study

We aim to evaluate the efficacy of combination therapy (enzalutamide plus abiraterone) relative to abiraterone monotherapy among patients with DNA damage response (DDR) mutations. The Alliance A031201 phase III trial (Morris et al., 2023) compares enzalutamide monotherapy with the combination therapy for metastatic castration-resistant prostate cancer (mCRPC) but lacks an abiraterone-monotherapy arm; therefore, we rely on external control data to estimate this contrast. The analysis analyze $N = 80$ combination-therapy patients from the Alliance trial alongside external abiraterone-treated control patients drawn from three sources: two single-arm clinical trials (AbiRace George et al. (2021), $N = 100$; PANTHER George et al. (2024), $N = 93$) and Duke patients who participated in the IRONMAN registry Mucci et al. (2022) ($N = 33$). Although the methodological development of our framework is presented primarily in terms of RWD as the source of external controls, single-arm trials administering the relevant SoC treatment can also provide external control data and can be accommodated through the same external-data component of the proposed model. Specifically, we applied the single-arm trial model described in Section 3.2.2, adapted to the four data sources in this analysis as

$$h(i,s,d) = \begin{cases} \exp\left(\zeta_s + x_{i_{s,d}}\gamma_{s,d}\right) \text{ if } d = \text{SAT1} \\ \exp\left(\zeta_s + \zeta_{s,d}\right) \text{ if } d \in \{\text{RWD1}, \text{RWD2}\} \\ \exp\left(\zeta_s - \zeta_{s,\text{RWD1}} - \zeta_{s,\text{RWD2}}\right) \text{ if } d = \text{RWD3}, \end{cases}$$

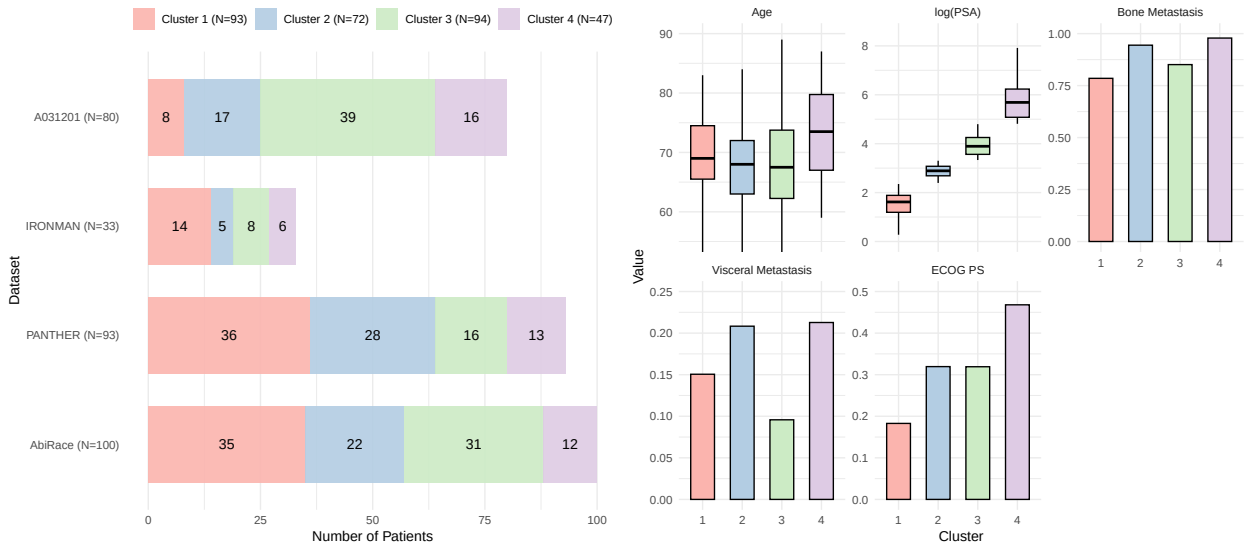


Figure 2: Clustering results. Left panel: Patient counts across estimated clusters by dataset. Each stacked bar represents a dataset, with colors indicating cluster membership; numbers within the bars denote the number of patients assigned to each cluster. Right panel: Baseline patient characteristics across estimated clusters. Boxplots summarize age and log-transformed PSA, whereas bars show the proportions of patients with bone metastasis, visceral metastasis, and ECOG performance status by cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
HRs	$2.47(0.91, 5.24)$	$3.52(1.13, 9.38)$	$1.75(1.05, 2.86)$	$1.28(0.64, 2.25)$
$\Delta_{\text{RMST}_{s,\text{SAT1}}}(12)$	$-0.82(-1.96, 0.06)$	$-0.82(-1.63, -0.09)$	$-0.82(-1.65, -0.07)$	$-0.45(-1.87, 0.93)$
$\Delta_{\text{RMST}_{s,\text{SAT1}}}(24)$	$-2.86(-6.49, 0.24)$	$-2.95(-5.71, -0.33)$	$-2.65(-5.26, -0.23)$	$-1.28(-5.20, 2.73)$
$\Delta_{\text{RMST}_{s,\text{SAT1}}}(36)$	$-5.61(-12.20, 0.49)$	$-5.98(-11.37, -0.68)$	$-4.84(-9.55, -0.43)$	$-2.08(-8.44, 4.63)$

Table 5: Posterior point and interval estimates for cluster-specific hazard ratios (HRs) and differences in RMST ($\Delta_{\text{RMST}_{s,\text{SAT1}}}(\tau)$) between combination therapy and abiraterone monotherapy at $\tau = 12$, 24, and 36 months.

where SAT1 corresponds to A031201, RWD1 to AbiRace, RWD2 to PANTHER, and RWD3 to IRONMAN.

Before fitting the hierarchical SAT model, we first performed clustering using five baseline covariates that are established prognostic factors in Halabi et al. (2014): age (discrete), prostate-specific antigen (PSA) (continuous), presence of bone metastasis (binary), presence of visceral metastasis (binary), and ECOG performance status (binary). Missing values were imputed in R using the `mice()` function with 50 imputations; for numeric variables, the mean across imputations was used, whereas for binary variables, the mode was used. For the QGW prior, we set $M = \lfloor \sqrt{n/2} \rfloor$ as described in Section 3.1.1. Figure 2 summarizes the clustering results, showing the cluster composition within each dataset and the baseline characteristics of each estimated cluster. The clustering structure was unchanged when M was halved or doubled, with pairwise adjusted Rand indices equal to 1. Noticeable differences in baseline characteristics are observed across the estimated clusters.

We then fitted the SAT model using the estimated cluster labels. The posterior mean of the

overall HR for the experimental arm, $\exp(\gamma)$, was 2.62 (95% credible interval: $0.27, 10.33$), providing no evidence of an overall survival benefit for the combination therapy compared with abiraterone monotherapy. Furthermore, the cluster-specific hazard ratios reported in Table 5 do not support a benefit in any subgroup. Instead, the estimates for Clusters 2 and 3 suggest higher mortality hazards under the combination therapy. In particular, the estimate for Cluster 3 is more precise, with a narrower credible interval that excludes 1.

Table 5 also reports the difference in RMST ($\Delta_{\text{RMST}_{s,\text{SAT1}}}(\tau)$), the causal estimand introduced in Section 3.2.1, at prespecified horizons of $\tau = 12$, 24, and 36 months (1, 2, and 3 years). Consistent with the HR results, the estimated $\Delta_{\text{RMST}_{s,\text{SAT1}}}(\tau)$ values for Clusters 2 and 3 are negative, indicating lower RMST for the combination therapy than under abiraterone monotherapy over the corresponding time horizons.

These cluster-specific findings should nevertheless be interpreted cautiously because uncertainty from the clustering stage is not propagated into the second-stage credible intervals. Thus, the reported uncertainty may be underestimated, and these findings should be viewed as exploratory rather than definitive evidence of subgroup-specific treatment effects.

6 Conclusion

We have developed a novel two-stage Bayesian framework for integrative analyses of RCT and RWD that facilitates baseline comparability and considers potential outcome heterogeneity across data sources.

In the first stage, we employ a Bayesian mixture model for clustering with a new nonparametric prior, the QGW prior, which offers greater flexibility than the existing GW prior and helps reduce label-switching issues arising from the exchangeability of mixture components and the resulting non-identifiability of component labels. The stick-breaking construction of the QGW prior introduces a natural asymmetry in the mixture weights, providing a useful labeling mechanism through their ordering. This model-based approach promotes covariate comparability by clustering observations with similar covariate distributions.

In the second stage, we develop a Bayesian hierarchical model for estimating treatment contrasts while accounting for heterogeneity across both clusters and datasets. We also formulate RMST-based causal estimands under the proposed model within the potential outcomes framework. In addition, we provide an extension of the model to accommodate single-arm trials with external control data.

Simulation studies demonstrate the strong performance of both stages. First, the QGW prior improves cluster recovery compared with the GW prior, maintaining robustness under higher-dimensional covariates. Second, incorporating RWD enhances estimation efficiency for overall treatment contrasts—especially when concurrent control data are limited—while demonstrating favorable MSEs for cluster-specific effects compared with the propensity score-integrated power prior approach.

In the case study, we applied the proposed framework to patients with DDR mutations from the Alliance A031201 trial, using external abiraterone-treated controls to estimate the treatment contrast between enzalutamide plus abiraterone combination therapy and abiraterone monotherapy. The estimated overall hazard ratio provided no evidence of an overall survival benefit for the combination therapy, while cluster-specific estimates for two subgroups suggested higher mortality hazards. The corresponding RMST contrasts were consistent with these findings.

Several aspects of the first-stage model provide opportunities for further development. The

current mixture model assumes conditional independence among the variables in $\boldsymbol{z}$ within each mixture component, although in practice covariates may exhibit dependence. Dependence among continuous covariates could be incorporated by using a multivariate normal kernel whose mean and covariance matrix serve as mixing parameters. More generally, dependence across variables of different types could be handled through latent Gaussian (Gaussian copula) modeling approaches (e.g., Pitt et al., 2006; Murray et al., 2013), in which latent variables are linked to the observed covariates via monotone or threshold mappings (e.g., Albert and Chib, 1993; Chib and Winkelmann, 2001).

The present work sets $a = b = 1$ for the stick-breaking fractions v_l in (2), a practical choice that facilitates posterior computation through closed-form full conditional distributions in the first-stage mixture model. Under more general $\text{Be}(a, b)$ specifications, Metropolis-Hastings algorithms could be used for posterior updates, although computation may become inefficient when a large number of clusters is anticipated, potentially resulting in slow convergence and poor mixing. More efficient posterior sampling algorithms would permit broader specifications of the QGW prior and facilitate systematic investigation of its sensitivity to the beta parameters. In addition, although the ordered-weight construction helps reduce label switching, it does not eliminate the problem entirely. When two or more components have similar weights, the ordering constraint may provide limited separation between their labels, and the corresponding components may remain difficult to distinguish, consistent with the identifiability considerations discussed by Stephens (2000). A fuller theoretical characterization of the support properties of the QGW prior is another important direction for future work.

Several extensions of the second-stage model are also possible. We focus on time-to-event endpoints and propose a Cox-based hierarchical model with tractable inference enabled by Gibbs sampling with closed-form full conditionals. The proposed modeling framework is not limited to survival outcomes and can be extended to other outcome types, such as continuous, discrete, or binary, under appropriate distributional assumptions (e.g., normal, negative binomial, and Bernoulli). In these cases, closed-form full conditionals may still be available, particularly through the use of Pólya–Gamma data augmentation for discrete and binary outcomes.

A further limitation arises from the separation of the clustering and outcome-analysis stages. The estimated cluster assignments from the first stage are treated as fixed inputs in the second stage, so uncertainty in clustering is not propagated into posterior inference for the second-stage parameters. Misclassification may alter the composition and number of retained clusters, thereby affecting the estimation of cluster-specific parameters and, in turn, their higher-level parameters. A fully joint Bayesian formulation could propagate uncertainty in cluster assignments into treatment-effect estimation. However, such a formulation would allow outcome information to influence the clustering, whereas our two-stage approach intentionally defines population comparability using pretreatment characteristics only. Developing formal approaches for propagating clustering-stage uncertainty while preserving this separation, as well as establishing general posterior-consistency results for the complete two-stage procedure, are important directions for future theoretical work.

Finally, because clustering is based on observed covariates, residual differences between the RCT and RWD populations due to unmeasured prognostic factors may remain. As with other approaches that integrate RCT and RWD, the validity of the proposed framework therefore depends on the assumption that important prognostic factors have been adequately measured and incorporated into the clustering and outcome models. Although the adaptive borrowing mechanism can reduce the influence of residual incomparability across data sources, unmeasured differences may still introduce

bias. Accordingly, the proposed approach should be viewed as mitigating, rather than eliminating, the impact of residual confounding.

Funding

This research was supported in part by the National Institutes of Health Grants R01 LM013352, R01 CA256157, and R01 CA249279, the United States Army Medical Research Awards HT9425-23-1-0393 and HT9425-25-1-0623, and the Prostate Cancer Foundation.

References

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Charles Bouveyron, Gilles Celeux, T Brendan Murphy, and Adrian E Raftery. *Model-based Clustering and Classification for Data Science: with Applications in R*, volume 50. Cambridge University Press, 2019.

Wei-Chen Chen, Chenguang Wang, Heng Li, Nelson Lu, Ram Tiwari, Yunling Xu, and Lilly Q Yue. Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. *Journal of Biopharmaceutical Statistics*, 30(3):508–520, 2020.

Siddhartha Chib and Rainer Winkelmann. Markov chain monte carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19(4):428–435, 2001.

Paul Damien, Petros Dellaportas, Nicholas G Polson, and David A Stephens. *Bayesian theory and applications*. OUP Oxford, 2013.

Chris Fraley and Adrian E Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

Sylvia Fruhwirth-Schnatter, Gilles Celeux, and Christian P Robert. *Handbook of Mixture Analysis*. CRC press, 2019.

Daniel J George, Susan Halabi, Elisabeth I Heath, A Oliver Sartor, Guru P Sonpavde, Devika Das, Rhonda L Bitting, William Berry, Patrick Healy, Monika Anand, et al. A prospective trial of abiraterone acetate plus prednisone in black and white men with metastatic castrate-resistant prostate cancer. *Cancer*, 127(16):2954–2965, 2021.

Daniel J. George, Susan Halabi, Mark T. Fleming, Elisabeth I. Heath, Ronald Tutrone, Linda Sutton, Young E. Whang, Brian E. Lewis, Michael Sandon Humeniuk, Michael Roger Harrison, Julie Kephart, Julia Hurrelbrink, Julia Rasmussen, Kellie Shobe, Monika Anand, Marco Reyes-Martinez, Lauren Howard, Steven R. Patierno, Jennifer A Freedman, and Andrew J. Armstrong. Overall survival from panther: A multicenter trial of apalutamide, abiraterone acetate plus prednisone in black and white patients with metastatic castration-resistant prostate cancer (mcrpc). *Journal of Clinical Oncology*, 42(16_suppl):5072–5072, 2024.

Jens Hainmueller. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1):25–46, 2012.

Susan Halabi, Chen-Yen Lin, W Kevin Kelly, Karim S Fizazi, Judd W Moul, Ellen B Kaplan, Michael J Morris, and Eric J Small. Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *Journal of Clinical Oncology*, 32(7):671–677, 2014.

Miguel A Hernán. The hazards of hazard ratios. *Epidemiology*, 21(1):13–15, 2010.

Brian P Hobbs, Daniel J Sargent, and Bradley P Carlin. Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. *Bayesian Analysis*, 7(3):639, 2012.

Joseph G Ibrahim and Ming-Hui Chen. Power Prior Distributions for Regression Models. *Statistical Science*, pages 46–60, 2000.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Hemant Ishwaran and Mahmoud Zarepour. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 20(1):50 – 67, 2005. doi: 10.1214/088342305000000016.

Liyun Jiang, Lei Nie, and Ying Yuan. Elastic priors to dynamically borrow information from historical data in clinical trials. *Biometrics*, 79(1):49–60, 2023.

Junjing Lin, Margaret Gamalo-Siebers, and Ram Tiwari. Propensity-score-based priors for Bayesian augmented control design. *Pharmaceutical Statistics*, 18(2):223–238, 2019.

Li-An Lin, Yafei Zhang, Walter Straus, and William Wang. Integrative Analysis of Randomized Clinical Trial and Observational Study Data to Inform Post-marketing Safety Decision-Making. *Therapeutic Innovation & Regulatory Science*, 56(3):423–432, 2022.

Fang Liu and Demosthenes Panagiotakos. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*, 22(1):287, 2022.

Wei-En Lu and Ai Ni. Causal effect estimation on restricted mean survival time under case-cohort design via propensity score stratification. *Lifetime Data Analysis*, 31(4):898–931, 2025.

Michael J Morris, Glenn Heller, David W Hillman, Olivia Bobek, Charles Ryan, Emmanuel S Antonarakis, Alan H Bryce, Olwen Hahn, Himisha Beltran, Andrew J Armstrong, et al. Randomized phase iii study of enzalutamide compared with enzalutamide plus abiraterone for metastatic castration-resistant prostate cancer (alliance a031201 trial). *Journal of Clinical Oncology*, 41(18): 3352–3362, 2023.

Lorelei A Mucci, Jacob Vinson, Theresa Gold, Travis Gerke, Julie Filipenko, Rebecca M Green, Simon G Anderson, Simone Badal, Anders Bjartell, Kim N Chi, et al. Ironman: A novel international registry of men with advanced prostate cancer. *JCO Global Oncology*, 8:e2200154, 2022.

Jared S Murray, David B Dunson, Lawrence Carin, and Joseph E Lucas. Bayesian gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665, 2013.

Ai Ni, Zihan Lin, and Bo Lu. Stratified restricted mean survival time model for marginal causal effect in observational survival data. *Annals of Epidemiology*, 64:149–154, 2021.

Panagiotis Papastamoulis. label. switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs. *Journal of Statistical Software*, 69:1–24, 2016.

Michael Pitt, David Chan, and Robert Kohn. Efefficient bayesian inference for gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006.

Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

Jayaram Sethuraman. A CONSTRUCTIVE DEFINITION OF DIRICHLET PRIORS. *Statistica Sinica*, pages 639–650, 1994.

Mingyang Shan, Douglas Faries, Andy Dang, Xiang Zhang, Zhanglin Cui, and Kristin M Sheffield. A Simulation-Based Evaluation of Statistical Methods for Hybrid Real-World Control Arms in Clinical Trials. *Statistics in Biosciences*, 14(2):259–284, 2022.

Matthew Sperrin, Thomas Jaki, and Ernst Wit. Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20:357–366, 2010.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

Monica Tang, Sallie-Anne Pearson, Robert J Simes, and Boon H Chua. Harnessing Real-World Evidence to Advance Cancer Research. *Current Oncology*, 30(2):1844–1859, 2023.

Hajime Uno, Brian Claggett, Lu Tian, Eisuke Inoue, Paul Gallo, Toshio Miyata, Deborah Schrag, Masahiro Takeuchi, Yoshiaki Uyama, Lihui Zhao, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32 (22):2380–2385, 2014.

Shailja Vaghela, Kaniz Afroz Tanni, Geetanjoli Banerjee, and Vanja Sikirica. A systematic review of real-world evidence (RWE) supportive of new drug and biologic license application approvals in rare diseases. *Orphanet Journal of Rare Diseases*, 19(1):117, 2024.

Sara Wade. Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A*, 381(2247): 20220149, 2023.

Chenguang Wang, Heng Li, Wei-Chen Chen, Nelson Lu, Ram Tiwari, Yunling Xu, and Lilly Q Yue. Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *Journal of Biopharmaceutical Statistics*, 29(5):731–748, 2019.

Chenguang Wang, Nelson Lu, Wei-Chen Chen, Heng Li, Ram Tiwari, Yunling Xu, and Lilly Q Yue. Propensity score-integrated composite likelihood approach for incorporating real-world evidence in single-arm clinical studies. *Journal of Biopharmaceutical Statistics*, 30(3):495–507, 2020.

Peng Yang, Yuansong Zhao, Lei Nie, Jonathon Vallejo, and Ying Yuan. SAM: Self-adapting mixture prior to dynamically borrow information from historical data in clinical trials. *Biometrics*, 79(4): 2857–2868, 2023.