Biometrics, 2024, 80(4), ujae136
https://doi.org/10.1093/biomtc/ujae136
Biometric Methodology

A multivariate Polya tree model for meta-analysis with event-time distributions

Giovanni Poli [1,*], Elena Fountzilas [2], Apostolia-Maria Tsimeridou[3], Peter Müller [4]

[1]Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Florence, 50134, Italy, [2]Department of Medical Oncology, St Luke's Clinic, Thessalonik, 55236, Greece, [3]Department of Investigational Cancer Therapeutics, University of Texas MD Anderson Cancer Center, Houston, TX 77030, United States, [4]Department of Statistics and Data Science, University of Texas at Austin, Austin, TX 78705, United States

[*]Corresponding author: Giovanni Poli, Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Florence 50134, Italy (giovanni.poli@unifi.it).

ABSTRACT

We develop a nonparametric Bayesian prior for a family of random probability measures by extending the Polya tree (PT) prior to a joint prior for a set of probability measures $G_1, \ldots, G_n$, suitable for meta-analysis with event-time outcomes. In the application to meta-analysis, G_i is the event-time distribution specific to study i. The proposed model defines a regression on study-specific covariates by introducing increased correlation for any pair of studies with similar characteristics. The desired multivariate PT model is constructed by introducing a hierarchical prior on the conditional splitting probabilities in the PT construction for each of the G_i. The hierarchical prior replaces the independent beta priors for the splitting probability in the PT construction with a Gaussian process prior for corresponding (logit) splitting probabilities across all studies. The Gaussian process is indexed by study-specific covariates, introducing the desired dependence with increased correlation for similar studies. The main feature of the proposed construction is (conditionally) conjugate posterior updating with commonly reported inference summaries for event-time data. The construction is motivated by a meta-analysis over cancer immunotherapy studies.

KEYWORDS: Gaussian process; nonparametric inference; survival analysis.

1 INTRODUCTION

We introduce a multivariate Polya tree (PT) model for inference on a set of dependent random distributions $\{G_i\ i = 1, \ldots, n\}$, suitable for meta-analysis of event-time data over multiple studies—or cohorts—in the motivating application. The motivating application is a meta-analysis over n cohorts in S studies, with each study consisting of multiple patient cohorts (ie, $n > S$) and G_i being the distribution of progression-free survival (PFS) for patients in cohort i. The level of dependence across G_i is modeled as a function of cohort-specific covariate vectors $\boldsymbol{x}_i$, including tumor type, treatment agent, study indicator, biomarker status, and more. We model the dependence between cohort-specific event-time distributions G_i by introducing a Gaussian process (GP) prior on the logit conditional splitting probabilities in the PT construction. We argue that a PT prior is a natural model for meta-analysis with event-time outcomes, which typically report a point estimate m_i for the median event time and a corresponding confidence interval (ℓ_i, h_i). We show that the triple $\boldsymbol{s}_i = (\ell_i, m_i, h_i)^\top$ and the sample size N_i (under some assumptions) are equivalent to reporting counts for the four intervals defined by ℓ_i, m_i and h_i. An appropriately defined PT prior for such data allows for easy posterior updating, greatly facilitating inference. In a very natural and principled way, the proposed model formalizes the integration of different sources of knowledge, including data and clinical expert information. Conditioning on the data is implemented through posterior updating of G_i, while expert knowledge about the similarity of different cohorts is encoded in the GP covariance function.

Several extensions of the PT models to priors for families of random probability measures have been proposed in the literature. In the upcoming discussion, we will refer to such models generically as multivariate PT. Some of the earlier references address the closely related problem of constructing PT priors for a multivariate distributions, that is, by way of splitting probabilities for multivariate intervals. This could in principle be used to define a family of random probability measures by way of the implied univariate marginals. For example, Yang et al. (2008) use such PT's to define a prior on a bivariate sample space. Jara et al. (2009) use similar models as a nonparametric prior for random effects in a semi-parametric regression. However, this approach is only practicable for a small number of random probability measures. A more general approach is proposed by Christensen and Ma (2019), who generate dependent random probability measures by adding an additional level in a hierarchical model, with a nonparametric hyper-prior on the common base measure for multiple PT's. Another approach proposed in Trippa et al. (2011) introduces a gamma process indexed by covariates. Ratios of probabilities under the gamma process define marginally beta-distributed splitting probabilities for dependent PT priors with desired correlation across multiple cohorts arising from

Received: December 12, 2023; Revised: October 3, 2024; Accepted: October 29, 2024

using the same underlying gamma process. Specifically in the context of meta-analysis, Branscum and Hanson (2008) developed a Polya tree mixture model for the random effects prior. Diana et al. (2023) introduce the replicate PT framework, which models correlation by imposing constraints on the parameters and replicating parts of the trees. In the approach that we propose in this article, we introduce correlations using the covariance function of a GP, directly modeling the correlation between splitting probabilities that define the cohort-specific event-time distributions G_i. The approach is most similar to the general dependent tail-free process (DTFP) that is defined in Jara and Hanson (2011), who then proceed to propose and implement the special case of the linear dependent tail-free process (LDTFP). The LDTFP uses a normal linear regression for the logit splitting probabilities. Flores and Müller (2024) build a model for meta-analysis using a nonparametric mixture of LDTFP's. Our intermediate model introduced in Section 3.2 is essentially another special case of DTFP with a tailored prior that allows borrowing information among heterogeneous cohorts.

The proposed construction is motivated by a meta-analysis over 174 published studies on early-phase cancer immunotherapy. Immunotherapy has shown promising efficacy results in several types of cancer. However, only a subgroup of patients benefit from this treatment, possibly due to patient and tumor heterogeneity. Depending on the tumor type, approximately 80% of patients do not respond or even develop hyper-progressive disease (hyper-progression), while a proportion of patients who initially responded eventually develop resistance. In addition, toxicity remains an issue with some patients developing serious immune-related adverse events. Finally, while the use of selected FDA-approved biomarkers is known to be associated with improved clinical outcomes in selecting patients receiving immunotherapy (Marabelle et al., 2020; Patel and Kurzrock, 2015), most immunotherapy trials are still conducted without biomarker selection. These considerations suggest that the use of robust predictive biomarkers (eg, gene expression or protein activation) could enable optimal therapy recommendations for patients with diverse tumor types. This requires the development of study designs that allow testing such hypotheses and provide inference on promising biomarkers. However, published studies are systematically underpowered to test hypotheses about biomarker subgroups. In this situation, meta-analysis, that is, the pooling of information across multiple studies, may provide useful.

Standard methods for meta-analysis are based on weighted linear regression with random effects (Schwarzer et al., 2015; Sutton and Abrams, 2001; Viechtbauer, 2010). See Ruberu et al. (2023) for an example of a recent application for cancer studies reporting relative risks, including a careful construction to accommodate different reporting modalities across studies. They implement meta-analysis using a parametric Bayesian inference model. Implementations of meta-analysis specifically for survival endpoints are discussed, for example, in Parmar et al. (1998) and Michiels et al. (2005). In particular, Michiels et al. (2005) discuss meta-analysis when only the median survival times are reported, including meta-analysis based on log median ratios across two conditions. This approach was used in Fountzilas et al. (2023b) to analyze same data as in our motivating application. We argue that such analysis fails to effectively model the heterogeneity of the data. An alternative to account for some of this heterogeneity could be the use of a parametric meta-regression model, but the small number of observations (especially for rare tumor types and less commonly used agents) limits the meaningful use of meta-regression. The proposed method introduces a practically feasible, fully nonparametric alternative in which information sharing across studies is established in a principled manner within the framework of an encompassing probability model.

2 A META-ANALYSIS OF CANCER IMMUNOTHERAPY STUDIES

We analyze data from a meta-analysis and systematic review of phase I/II clinical trials assessing the effect of biomarkers on clinical outcomes in patients with solid tumors (Fountzilas et al., 2023a). The full data is available from Fountzilas et al. (2023b). The analysis did not aim to demonstrate whether specific biomarkers are predictive of benefit from immunotherapy. Such an analysis would be of limited validity for rarely evaluated biomarkers. The goal was to determine whether, in general, the selection of patients based on biomarkers could be associated with clinical benefit. Data were collected using a PubMed search for phase I/II cancer clinical trials evaluating immune checkpoint inhibitors approved by FDA between 2018 and 2020. Only studies that reported summaries stratified by biomarker status were selected. In total, 174 clinical studies with a total of 19 178 patient responses were included in the analysis in Fountzilas et al. (2023b). Studies investigated several biomarkers, including PD-L1 expression (111 studies), tumor mutational burden (20 studies), and microsatellite instability/mismatch repair deficiency (10 studies).

In this analysis we focus on progression-free survival (PFS) as a particular endpoint, which is reported by $S = 33$ studies, for a total of $n = 84$ cohorts. Here, a cohort refers to a subset of patients in a study for which results are reported separately, including in particular marker-positive and -negative cohorts. However, some studies break down results by additional characteristics beyond biomarker status, thereby contributing with more than two cohorts. The reported summaries for PFS include a point estimate for the median PFS (m_i) and a corresponding confidence interval (ℓ_i, h_i) for each cohort, $i = 1, \ldots, n$. In the proposed inference approach, we model the unknown underlying distribution G_i of PFS that generated event times y_{ij} for N_i patients in cohort i. However, posterior updating can only condition on the available summaries (m_i, ℓ_i, h_i) and the known sample size N_i. As in most meta-analyses, patient-level data y_{ij} are not available.

Let i^+ and i^- denote the marker-positive and marker-negative cohorts of the same study, that is, $\boldsymbol{x}_{i^+}$ and $\boldsymbol{x}_{i^-}$ differ only by biomarker status. Let G_i^+ and G_i^- denote the corresponding event-time distributions with medians M_i^+ and M_i^-. The main inference goal is the comparison of medians, ie, inference about the hypothesis $M_i^+ > M_i^-$, formalizing the motivating question about the use of biomarkers in cancer immunotherapy. The proposed multivariate PT model represents an attractive statistical inference approach in this context, as it allows evaluation of the likelihood function for (m_i, ℓ_i, h_i) and (conditionally)

conjugate posterior updating. As a side benefit, the borrowing of strength across cohorts significantly improves inference for rare conditions, and allows more precise inference, for example, for event-time distributions for patients with rare tumor types, or less commonly used treatment agents.

3 A MULTIVARIATE POLYA TREE FOR EVENT-TIME OUTCOMES

We first introduce notation by way of reviewing the construction of a univariate PT prior in Section 3.1. In Section 3.2, we review the construction of the DTFP of Jara and Hanson (2011), which extends the construction to a multivariate PT with common partitioning subsets, which is then finally, in Section 4, extended to allow for different partitioning subsets for each distribution. The latter is needed for the desired meta-analysis with event-time data.

3.1 Univariate Polya trees

The PT (Lavine, 1992; 1994) is a prior distribution for a random probability measure G defined over a sample space S. It is constructed recursively using nested partitions $\pi_d = \{B_{e_1\cdots e_d};\ e_\ell \in \{0,1\}, \ell = 1,\ldots,d\}$, $d = 1, 2, \ldots$, of the sample space, starting with $S = B_0 \cup B_1$ and recursively refining the partition with $B_{e_1\cdots e_d} = B_{e_1\cdots e_d 0} \cup B_{e_1\cdots e_d 1}$. That is, $B_{e_1\cdots e_d 0}$ and $B_{e_1\cdots e_d 1}$ are defined by splitting $B_{e_1\cdots e_d}$ into a left and right binary partitioning subset. Following standard notation, we use subscript 0 for left and 1 for right partitions. We use $\varepsilon_d = e_1\cdots e_d \in \{0,1\}^d$ to uniquely identify a partitioning subset $B_{\varepsilon_d} \in \pi_d$. A prior model on G is implicitly defined by a prior on the conditional splitting probabilities $Y_{\varepsilon_d 0} \equiv G(B_{\varepsilon_d 0} \mid B_{\varepsilon_d})$, together with the choice of the partitioning subsets B_{ε_d}. The standard PT prior assumes $Y_{\varepsilon_d 0} \sim \text{Be}(\alpha_{\varepsilon_d 0}, \alpha_{\varepsilon_d 1})$ (and $Y_{\varepsilon_d 1} = 1 - Y_{\varepsilon_d 0}$). The construction can be described as a sequence of increasingly refined random histograms, with bins defined by B_{ε_d} and corresponding probabilities $G(B_{\varepsilon_d}) = \prod_{\ell=1}^{d} Y_{e_1\cdots e_\ell}$, and is illustrated in Web Figure 5 in the online Supplementary Materials.

The construction defines a random probability measure $G \sim \text{PT}(\mathcal{A}, \Pi)$, where $\mathcal{A} = \{\alpha_{\varepsilon_d 0}, \alpha_{\varepsilon_d 1} :\ d = 1, 2, \ldots\}$ is the set of hyper-parameters that index the beta priors on $Y_{\varepsilon 0}$ and $\Pi = \{\pi_d :\ d = 1, 2, \ldots\}$ is the nested partition sequence. The hyperparameters $\mathcal{A}$ and Π can be chosen to ensure a desired prior mean, $\mathbb{E}[G] = G_0$. Expressing prior information by way of a centering distribution or prior mean is a common feature in nonparametric Bayesian models. In the PT model, there are two main strategies to achieve a desired prior centering. The first is to fix Π and adjust the hyperparameters $(\alpha_{\varepsilon_d 0}, \alpha_{\varepsilon_d 1})$ to ensure $\mathbb{E}[Y_{\varepsilon_d 0}] = G_0(B_{\varepsilon_d 0} \mid B_{\varepsilon_d})$. The alternative strategy is to fix $\mathcal{A}$ to ensure $\mathbb{E}[Y_{\varepsilon_d 0}] = 0.5$ and then achieve the desired prior centering by using dyadic quantiles under G_0 as partitioning subsets, that is; defining $B_{\varepsilon_d 0}$ such that $G_0(B_{\varepsilon_d 0} \mid B_{\varepsilon_d}) = 0.5$ (Lavine, 1992; 1994). A common choice for α_{ε_d} in the latter case is $\alpha_{\varepsilon_d 0} = \alpha_{\varepsilon_d 1} = c \cdot (d+1)^2$, which ensures a continuous random distribution under a PT prior. The hyperparameter c is a scalar precision parameter, which is widely discussed in the PT literature. We will later use both strategies. We use fixed partitioning subsets for levels $d = 1, 2$ (matching the intervals defined by ℓ_i, m_i and h_i), and dyadic splits for $d > 2$. PT priors have several attractive and useful properties, including conjugacy under i.i.d. sampling and flexibility in encoding prior beliefs. Compared to other nonparametric priors, the main drawback of the PT model is the lack of smoothness of the density of G. However, this is of less concern in survival analysis, as the primary target is often the cumulative density function or, equivalently, the survival function shown in the Kaplan–Meier curve.

3.2 Multivariate Polya tree with Gaussian process dependence

Jara and Hanson (2011) introduce a prior for a family of random probability measures $\{G_i;\ i = 1,\ldots,I\}$ based on a generalization of a univariate PT prior. We review their construction, introducing some variations in anticipation of the next extension. In all variations, G_i remains a random probability measure over a sample space S with (at least approximately) a marginal PT prior.

Recall that in the motivating application n is the number of cohorts with available data on PFS. In anticipation of posterior predictive inference for future studies, we are setting up the model for $I > n$ cohorts, including cohorts indexed by $i \in \{n+1,\ldots,I\}$ without observed data. We first set up a model sharing a *common partitioning sequence* Π across all i. Let then $Y^{(i)}_{\varepsilon_d 0} = G_i(B_{\varepsilon_d 0} \mid B_{\varepsilon_d})$ denote the splitting probabilities under G_i and let $\eta(p) = \log\{p\ (1-p)^{-1}\}$ denote a logistic link function. The model maintains independence of splitting probabilities $Y^{(i)}_{\varepsilon}$ for different partitioning subsets B_ε within the same tree, but allows the splitting probabilities for the same ε to be correlated across i. This is achieved by introducing a Gaussian process (GP) prior on $Z^{(i)}_{\varepsilon 0} = \eta(Y^{(i)}_{\varepsilon_d 0})$. The GP is indexed by cohort-specific covariates $\boldsymbol{x}_i$ and replaces the independent beta prior of the univariate PT construction. That is, we assume $\{Z^{(i)}_{\varepsilon_d 0}\}_{\boldsymbol{x}_i \in X} \sim \text{GP}(\mu_{\varepsilon_d 0}, K_{\varepsilon_d 0})$ with mean function $\mu_{\varepsilon_d 0}$ and covariance function $K_{\varepsilon_d 0}$ (see below for $\mu_{\varepsilon_d 0}$ and $K_{\varepsilon_d 0}$). There is a separate, independent GP prior for each $\varepsilon_d 0 = \epsilon_1 \cdots \epsilon_d 0$ (and recall that any $Y^{(i)}_{\varepsilon_d 1}$ is implied as $1 - Y^{(i)}_{\varepsilon_d 0}$). Dependence is limited to partitioning subsets up to a certain depth D of the tree, with GP priors for each $\varepsilon_d 0$ up to level D, and independent beta priors beyond. We write $(G_1,\ldots,G_I) \sim \text{mvPT}_{GP}(\Pi, D, \mathcal{K}, \mathcal{A})$. The parameter $\mathcal{K}$ is the set of $2^D - 1$ pairs of mean and covariance functions $\left(\mu_{\varepsilon_d 0}(\cdot); K_{\varepsilon_d 0}(\cdot,\cdot)\right)$ that define the GP priors. The parameter Π is the (common) nested partition sequence. Finally, the set $\mathcal{A}$ is defined as in the univariate PT prior and collects the parameters $\alpha_{\varepsilon_d 0}, \alpha_{\varepsilon_d 1}$ for $d > D$. Posterior updating is similar to the univariate PT and depends on the counts of observations in each subinterval of Π. Logit splitting probabilities $Z^{(i)}_{\varepsilon_d 0}$ associated with the same sequence $\varepsilon_d 0$ are dependent across i (ie, cohorts) and it is convenient to sample them jointly. Using a logit link allows easy posterior updates using the Polya-Gamma sampler introduced by Polson et al. (2013). The same tree-based logit normal is introduced as logistic-tree normal in Wang et al. (2022) and LeBlanc and Ma (2022), where it is used as a prior for categorical probabilities in a mixed membership model.

For the hyperparameters, we propose choices that imply a desired marginal distribution for G_i, similar to the univariate PT. For details, see below (in the context of setting up marginal

moments $\mu_{\varepsilon_d 0}$ and $\sigma^2_{\varepsilon_d 0}$ for the GP prior). For the shared nested partition sequence Π we use the dyadic quantiles of a base measure G_0. Additionally, for $\varepsilon_d \in \{0,1\}^d$ at levels $d = 1, \ldots, D$ we need a covariance function $K_{\varepsilon_d 0}(\cdot,\cdot)$ and mean process $\mu_{\varepsilon_d 0}(\cdot)$ for the GP prior. We factor the covariance function as $K_{\varepsilon_d 0}(\boldsymbol{x},\boldsymbol{x}') = \sigma^2_{\varepsilon_d 0} \cdot R(\boldsymbol{x},\boldsymbol{x}')$, and first consider the moments of the marginal distribution $Z^{(i)}_{\varepsilon_d 0} \sim N(\mu_{\varepsilon_d 0}, \sigma^2_{\varepsilon_d 0})$. Our choice is based on fixing the parameters $(\mu_{\varepsilon_d 0}, \sigma^2_{\varepsilon_d 0})$ to approximately match a $\text{Be}(\alpha_{\varepsilon_d 0}, \alpha_{\varepsilon_d 1})$ prior on $\eta^{-1}(Z^{(i)}_{\varepsilon_d 0})$, that is, the prior under a univariate PT. Although beta and logistic-normal distributions are never exactly equal, any beta distribution can be approximated with the logistic-normal distribution that minimizes Kullback-Leibler divergence (Aitchison and Shen, 1980). Let ψ and ψ' denote the digamma and trigamma functions, respectively. We use $\mu_{\varepsilon_d 0}(\boldsymbol{x}) = \psi(c \cdot [d+1]^2) - \psi(c \cdot [d+1]^2) = 0$ and $\sigma^2_{\varepsilon_d 0} = 2 \cdot \psi'(c \cdot [d+1]^2)$. Having specified the marginal moments $\mu_{\varepsilon_d 0}$ and $\sigma^2_{\varepsilon_d 0}$, we are left to specify $R(\boldsymbol{x},\boldsymbol{x}')$, ie, the correlation between logit-transformed splitting probabilities for cohorts with covariates $\boldsymbol{x}$ and $\boldsymbol{x}'$. The correlation function $R(\boldsymbol{x},\boldsymbol{x}')$ is used to introduce clinical expert judgment on similarity of the event-time distributions G_i. See Section 5 for an example of constructing $R(\boldsymbol{x},\boldsymbol{x}')$ tailored to our application. For levels $d > D$ we define partitioning subsets $B_{\varepsilon_d 0}$ and beta parameters $\alpha_{\varepsilon_d 0}$, $\alpha_{\varepsilon_d 1}$ to achieve a desired prior mean G_0 as described in Section 3.1, using $\alpha_{\varepsilon_d} = cd^2$. In summary, $R(\boldsymbol{x},\boldsymbol{x}')$ introduces clinical expert judgment on how similar event-time distributions for different cohorts are likely to be, and the precision parameter c and the centering measure G_0 fix prior uncertainty and expectation of the marginal prior on G_i. Defining prior elicitation for the GP parameters by approximating the beta prior in the standard PT construction allows us to use the same c and G_0 to characterize splitting probabilities across all levels, including $d = 1, 2$ with the logit-normal GP prior as well as $d > 2$ with the beta prior.

If desired, it is possible to define hyperpriors and potentially learn about parameters in the covariance functions (Murphy, 2012, Chapter 15). However, this possibility is not explored here. Finally, let $B \subset A$ denote any two nested subsets, and let $\boldsymbol{Z} = \{Z^{(i)}\}_{i=1}^I$ with $Z^{(i)} = \eta\{G_i(B \mid A)\}$. Via Monte Carlo prior simulation, the proposed model allows to evaluate $\mathbb{E}[Z^{(i)}]$ and $\mathbb{C}ov(Z^{(i)}, Z^{(i')})$ for any i, i'. We shall use this later. Pseudo code for this prior simulation is available online as Supplementary Materials. Figure 1 shows a random sample $(G_i, i = 1, \ldots, 20)$ from a mvPT for two different correlation matrices $R(\boldsymbol{x},\boldsymbol{x}')$, and illustrates how two marginal random distributions are constructed from $G_i(B^{(i)}_{\varepsilon_d 0} \mid B^{(i)}_{\varepsilon_d})$.

4 A POLYA TREE PRIOR FOR META-ANALYSIS WITH EVENT-TIME DATA

4.1 Multivariate Polya tree with study-specific partitions

Recall the format of the data with $\boldsymbol{s}_i = (\ell_i, m_i, h_i)$ and sample size N_i for each cohort in the meta-analysis. We assume that m_i and (ℓ_i, h_i) were determined as the intersections of the Kaplan–Meier (KM) survival curve (Kaplan and Meier, 1958) and the corresponding error bounds, respectively, with the 0.5 threshold. We assume that the error bounds are based on the KM estimator and the Greenwood formula (Greenwood, 1926; Hosmer et al., 2011). Conditioning on a censoring pattern (ie, the order in which observed event times and censoring events occur—see below about updating this assumption) $\boldsymbol{s}_i$ implies counts of observations in each of the four sub-intervals determined by (ℓ_i, m_i, h_i). If these subintervals match the partitioning subsets in the first two levels of the marginal PT construction, then the counts are a sufficient statistic for the posterior distribution of $Y^{(i)}_0, Y^{(i)}_{00}, Y^{(i)}_{10}, i = 1, \ldots, n$, ie, to update knowledge on G_i. We therefore replace the shared partition sequence Π of the $\text{mvPT}_{GP}(\Pi, D, \mathcal{K}, \mathcal{A})$ by a set of *cohort-specific partition sequences* $\{\Pi_i\}_{i=1}^I$, with $\pi^{(i)}_1 = \{[0, m_i), [m_i, +\infty)\}$ and $\pi^{(i)}_2 = \{[0, \ell_i), [\ell_i, m_i), [m_i, h_i), [h_i, +\infty)\}$. Nested partitions π_d at deeper levels $d > 2$ are constructed by dyadic splits of the parent set $B^{(i)}_{\varepsilon_{d-1}}$ such that $G_0(B^{(i)}_{\varepsilon_d} \mid B^{(i)}_{\varepsilon_{d-1}}) = 0.5$. We refer to the extended model as $\text{mvPT}_{GP}(\{\Pi_i\}_{i=1}^I, D, \mathcal{K}, \mathcal{A})$, with $\{\Pi_i\}_{i=1}^I$ replacing the common shared partitioning sequence Π of the earlier construction. The extension requires careful consideration of the mean process $\mu^{(i)}_{\varepsilon_d 0}$ and the covariance function $K^{(i)}_{\varepsilon_d 0}$. Note the added superindex for cohort i, to allow for different $B^{(i)}_{\varepsilon_d}$. The elicitation involves expected values and covariances for each (logit transformed) conditional probability $G_i(B^{(i)}_{\varepsilon_d 0} \mid B^{(i)}_{\varepsilon_d})$ and $G_{i'}(B^{(i')}_{\varepsilon_d 0} \mid B^{(i')}_{\varepsilon_d})$, which now refer to possibly very different sets $B^{(i)}_{\varepsilon_d 0}$ and $B^{(i)}_{\varepsilon_d 0}$. A principled and coherent specification of such quantities is challenging. We use the following construction to reduce the problem to the earlier case of shared Π. We first consider a process with shared partitions $\mathcal{G}^* = (G^*_i, i = 1, \ldots, I) \sim \text{mvPT}_{GP}(\Pi, D, \mathcal{K}, \mathcal{A})$, defined as in the previous section.

Under $\mathcal{G}^*$ we can then by the earlier discussed prior simulation evaluate probabilities for any events. In particular, we can evaluate expected values and covariances for logit conditional probabilities $G_i(B^{(i)}_{\varepsilon_d 0} \mid B^{(i)}_{\varepsilon_d})$, as needed for the construction of $\text{mvPT}_{GP}(\{\Pi_i\}_{i=1}^I, D, \mathcal{K}, \mathcal{A})$.

Finally, we have to select partition sequences Π_i for future cohorts $(i > n)$. We proceed with ℓ_i defined as the median of ℓ_k, $k = 1, \ldots, n$, and similarly for m_i and h_i.

4.2 Censoring patterns and posterior inference

One feature of the proposed model is that it allows to condition on all available information, beyond only the median point estimates. Instead, we condition on the entire reported triple $\boldsymbol{s}_i$, including confidence intervals (if available). To map $\boldsymbol{s}_i$ to counts used for posterior updating of PT parameters, we need to make assumptions about the censoring pattern (ie, the sequence of observed and censored event times). We start by assuming a distribution for censoring times, $C^{(i)}_j \sim H$ for patient j in study (cohort) i. To update an assumed censoring pattern, we employ an ABC-like (Marin et al., 2012) Metropolis-Hastings scheme. Specifically, we start with an initial assumption for the censoring pattern and then update following Metropolis-Hastings transition probabilities, accepting only transitions that accurately reproduce the observed triple (ℓ_i, m_i, h_i). The latter is assumed to be derived from a Kaplan–Meier estimator. We refer to the simulation as an ABC algorithm since we generate new values for $C^{(i)}_j$ and $T^{(i)}_j$ accepting only those that imply a match with the

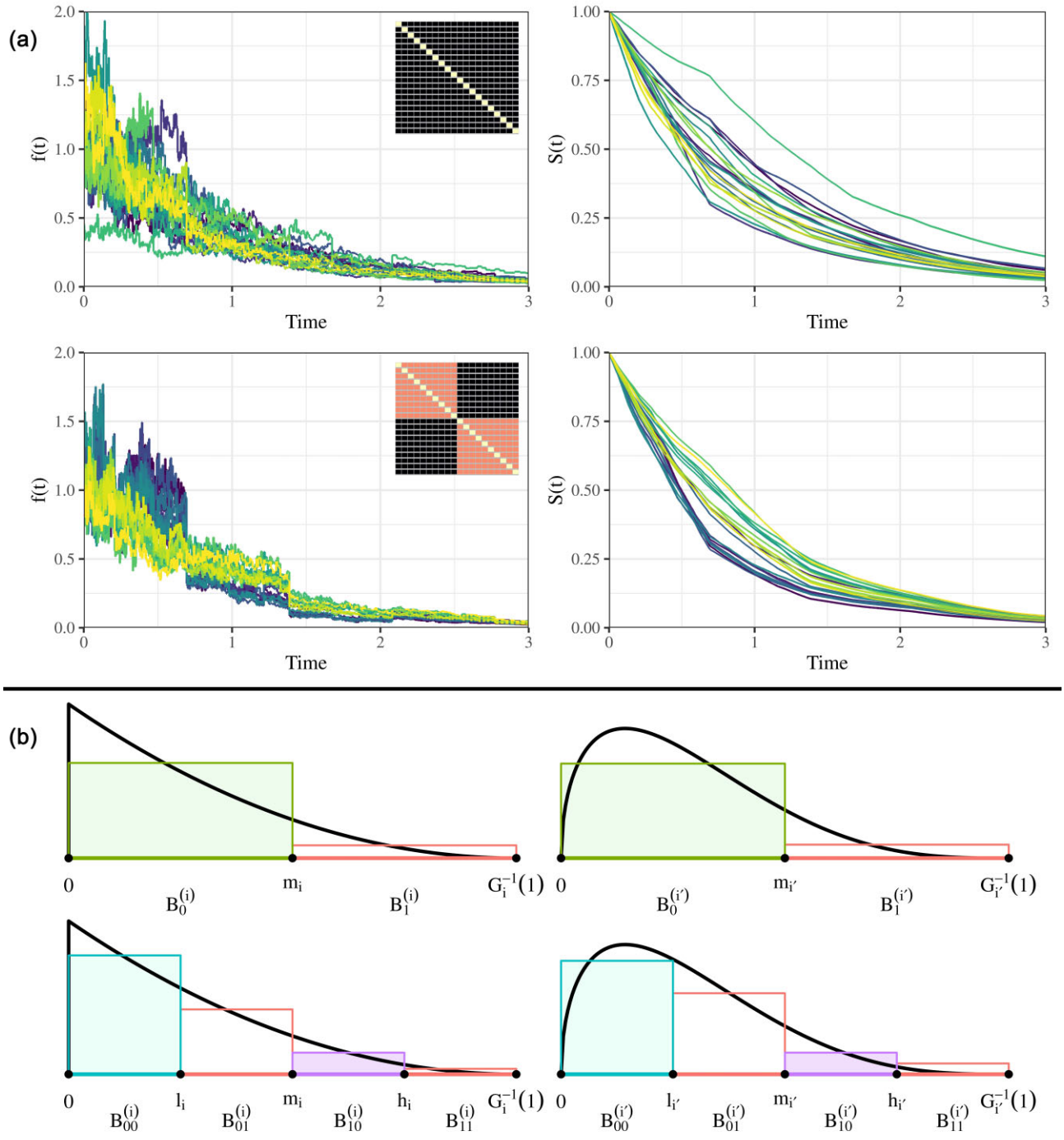

FIGURE 1 The top four panels (box a) show random samples $(G_1, \ldots, G_n)$ from a mvPT prior, using two different choices of $R(\boldsymbol{x}, \boldsymbol{x}')$ (plotted in the square insert in the left panels). The left panels show the densities. The right panels show the survival functions. Both use $G_0 = \text{Exp}(x \mid \lambda = 1)$ and precision parameter $c = 5$. For the first row, $R(\boldsymbol{x}, \boldsymbol{x}')$ is the identity (implying in particular that G_i are exchangeable); in the second row it is block diagonal (the G_i are partially exchangeable within each block, for example , marker-positive and -negative cohorts). The bottom panels (box b) show the random probabilities $G_i(B_{\varepsilon_d}^{(i)})$ for $d = 1, 2$ (in the bottom two rows) for $i \neq i'$ (in the two columns). Splitting probabilities $Y_{\varepsilon_d}^{(i)}$ for subsets $B_{\varepsilon_d}^{(i)}$ marked with the same color are dependent (across i), unfilled bins are deterministic given the other bins (ie, $Y_{\varepsilon_d 0}^{(i)}$ implies $Y_{\varepsilon_d 1}^{(i)}$). The overlaid density curve shows the density defined by the limit as $d \to \infty$.

reported statistics $\boldsymbol{s_i}$. See Appendix B for details of this simulation, including the initialization.

Keep in mind that this simulation is only imputing censoring patterns; there is no notion of the posterior simulation of parameters. This simulation-based approach enables us to accommodate different levels of information provided for each study. For example, for a study that reports no censoring events, we use deterministic counts, while a study that reports the number of censoring events can be treated differently from a study that reports no details on censoring. Finally, for a study that reports confidence intervals for different coverage probability than others, it is straightforward to account for this choice in the derivation of point estimate and confidence intervals for the median from the Kaplan–Meier plot.

4.3 Posterior summaries

Recall that G_i is the distribution of event times in cohort i, with covariates $\boldsymbol{x}_i$ and that the model is jointly defined on $G_1, \ldots, G_n$ ($i = 1, \ldots, n$ are the cohorts with observed data) and $G_{n+1}, \ldots, G_I$ (future cohorts). Let $\mathrm{Md}(P)$ be the median of a probability measure P, let $M_i = \mathrm{Md}(G_i)$ denote the median of G_i and let $\mathcal{S} = \{\boldsymbol{s_i}\}_{i=1}^n$ denote the observed data. We suggest reporting $p(M_i \mid \mathcal{S})$ to summarize inference under the proposed model. If point estimates are needed, we use the posterior median $\widehat{M}_i = \mathrm{Md}\big\{p(M_i|S)\big\}$. For the observed cohorts, $i = 1, \ldots, n$, the posterior distribution $p(M_i \mid \mathcal{S})$ summarizes updated knowledge. However, from an inferential point of view, the main interest is on the posterior distribution $p(M_i \mid \mathcal{S})$ for future cohorts $i = n + 1, \ldots, I$. Moreover, research often focuses on populations that correspond to multiple covariate vectors $\boldsymbol{x}_i$, that is; mixtures of multiple future cohorts i with $i \in A \subseteq \{n + 1, \ldots, I\}$. For example, inference on marker-positive patients is naturally represented as a mixture where A is the set of all marker-positive future cohorts. Event-time distributions for such populations are implicitly defined as $P = \sum_{i \in A} \pi_i \, G_i$, weighting different cohorts with possibly non-uniform weights π_i.

One aspect to keep in mind with inference on M_i, $i > n$, is that $p(M_i \mid \mathcal{S})$ also includes study-to-study variation. In contrast, inference under classical meta-regression usually reports P-values for fixed effects θ, that is, an average effect for future studies with particular characteristics $\boldsymbol{x}$. To define a comparable inference summary under the proposed nonparametric model, let $\theta = \{Y_{\varepsilon_d}^{(h)} : d = 1, \ldots \; ; \; h = 1, \ldots, n\}$ denote the conditional splitting probabilities for the observed studies and define $\bar{G}_i = \mathbb{E}[G_i \mid \theta]$ and $\bar{M}_i = \mathrm{Md}(\bar{G}_i)$, $i = n + 1, \ldots, I$. Then $\bar{M}_i$ is a function of θ and $p(\bar{M}_i \mid \mathcal{S})$ reports uncertainty on the future cohorts without cohort-to-cohort variation. In the upcoming results, we report summaries of $p(\bar{M}_i \mid \mathcal{S})$ or credible intervals based on it as alternatives to $p(M_i \mid \mathcal{S})$.

Recall the definition of G_i^+, G_i^-, M_i^+ and M_i^- as the event-time distributions and corresponding medians for matching marker-positive and marker-negative cohorts. Similarly, let P^+ and P^- and corresponding medians M_P^+, M_P^- refer to mixture populations differing only by the presence of biomarkers. Inference summaries reported in the upcoming discussion will focus on the paired comparison of M_i^+ vs. M_i^- and M_P^+ vs. M_P^-. Alternatively, we will also report similar quantities for the expected values, $\bar{M}_i{}^+$ vs. $\bar{M}_i{}^-$ and $\bar{M}_P^+$ vs. $\bar{M}_P^-$.

5 CORRELATION FUNCTION AND PRIOR SPECIFICATION

We describe the elicitation of the mvPT$_{GP}$ prior that is used for the motivating meta-analysis problem and in the simulations, with minor adjustments. In the studies under consideration, median PFS is usually reported for within the first few months. We therefore use a half-Cauchy with $\sigma = 3.5$ (in months) as centering distribution G_0 to allow a heavy right tail. Next, we fix c at a weakly informative value $c = 5$, which implies an *a priori* 95% credible interval of $[1.08; 11.52]$ on the median for a new study G_i.

We construct the correlation function $R(\boldsymbol{x}, \boldsymbol{x}')$ to represent clinical judgment about the level of similarity between the event-time distributions for any pair of cohorts i, i'. We first describe the construction of $R(\boldsymbol{x}, \boldsymbol{x}')$ for any two cohorts with matching biomarker status. We proceed by introducing an additive similarity score, adding points for each matching categorical covariate $x_j = x_j'$, which is then rescaled to a unit maximum. The covariates that are expected to have the strongest association with PFS are tumor type and agent. If cohorts i and i' share either of these two characteristics, we record an increment of 2 points each for the similarity score. Tumor type includes "*other*." Two cohorts with "*other*" as the tumor type are considered unmatched. Covariates that are judged to be less likely strongly associated with PFS include biomarkers status, study phase (1, 1/2, or 2), line of therapy (1, $\geq$ 2, or any), and type of therapy (combination, combination-or-monotherapy, or monotherapy). For each matching secondary covariate, we record a 0.5 increment of the similarity score. Next, 1 point was added for cohorts i, and i' that are part of the same study, and an extra point is added on the diagonal for $i = i'$ to add a nugget in the implied covariance function. For cohorts i, i' with different biomarker status, we only apply the rule about the shared study, recording a similarity score of 1 point for cohorts within the same study. The reduced correlation for cohorts with different biomarker status avoids oversmoothing of biomarker effects.

The described construction implies a maximum similarity score of 8. Rescaling to a maximum of 1 defines $R(\boldsymbol{x}, \boldsymbol{x}')$. In summary, letting ω_j denote the previously introduced covariate-specific weights, and assuming that x_{i1} is a study indicator, and x_{i2} is an indicator for biomarker status, we define $R(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$ for $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ij})$ and $\boldsymbol{x}_{i'} = (x_{i'1}, \ldots, x_{i'J})$ as follows:

$$R(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \frac{\mathbb{I}(i = i') + \omega_1 \cdot \mathbb{I}(x_{i1} = x_{i'1}) + \mathbb{I}(x_{i2} = x_{i'2}) \cdot \sum_{j=2}^{J} \omega_j \cdot \mathbb{I}(x_{ij} = x_{i'j})}{1 + \sum_j \omega_j}. \tag{1}$$

TABLE 1 Simulation setup.

	Tumor Type 1 $\mathbf{Exp(\lambda)}$	**Tumor Type 2** $\mathbf{\mathcal{HN}(\lambda)}$	**Tumor Type 3** $\mathbf{0.5\,Exp(\lambda) + 0.5\,\mathcal{HN}(\lambda)}$
A0	$\lambda_i^+ = (2.5 + 0.5) \cdot \alpha_i$ $\lambda_i^- = 2.5 \cdot \alpha_i$ $S_i = 5$	$\lambda_i^+ = (3 + 0.5) \cdot \alpha_i$ $\lambda_i^- = 3 \cdot \alpha_i$ $S_i = 5$	$\lambda_i^+ = (3.5 + 0.5) \cdot \alpha_i$ $\lambda_i^- = 3.5 \cdot \alpha_i$ $S_i = 3$
A1	$\lambda_i^+ = (2.5 + 0.5 + 1) \cdot \alpha_i$ $\lambda_i^- = (2.5 + 1) \cdot \alpha_i$ $S_i = 5$	$\lambda_i^+ = (3 + 0.5 + 1) \cdot \alpha_i$ $\lambda_i^- = (3 + 1) \cdot \alpha_i$ $S_i = 5$	$\lambda_i^+ = (3.5 + 0.5 + 1) \cdot \alpha_i$ $\lambda_i^- = (3.5 + 1) \cdot \alpha_i$ $S_i = 2$

The table shows the assumed cohort-specific medians. Each event-time distributions are parameterized in terms of the median λ. Parameters $\alpha_i \sim \mathcal{U}(0.8, 1.2)$ are study-specific multiplicative effect on the median. S_i reports the number of simulzated studies sampled with each distribution.

See Appendix C for an argument that R defines a positive semi-definite correlation matrix, and for suggested adjustments if a similar construction does not define a positive definite matrix. In our application, we apply (1) only for categorical covariates, but a similar kernel construction can be adapted for continuous variables. Also, keep in mind that $R(\boldsymbol{x}, \boldsymbol{x}')$ is only used to construct a valid $I \times I$ covariance matrix for the given studies. The emphasis is on constructing a suitable $R(\boldsymbol{x}, \boldsymbol{x}')$ to represent prior judgment.

The level of censoring for PFS is only moderate. We assume an exponential distribution with mean 10 for the censoring distribution in all cohorts.

6 SIMULATION STUDY

The simulations aim to assess inference in a realistic scenario that calls for standard meta-analysis, and to show that the assumed dependence across cohorts can compensate for the more restrictive borrowing of strength under a parametric model. We construct a simulation truth to mimic the setup in the motivating study, using the Kaplan–Meier estimator to generate simulated datasets of summaries $\boldsymbol{s}_i = (\ell_i, m_i, h_i)$ and for each hypothetical dataset, we compare inference under the mvPT model versus standard methods for meta-analysis.

We assume three tumor types $(TT1, TT2, TT3)$, and two agents $(A0, A1)$, with $A1$ being associated with higher median PFS. The resulting six cases were further divided on the basis of biomarker status (positive or negative), resulting in 12 types of cohorts shown in Table 1. Biomarker-positive status is assumed to add a positive offset of 0.5 to the median. Simulating a certain number of studies (see Table 1) for each of the combinations of tumor and agent, we generated a total of 25 hypothetical studies, with a marker-positive and a marker-negative cohort for each study, resulting in a total of $n = 50$ cohorts for each simulated dataset. Finally, for each study, we generate a study-specific multiplicative random effect on the median, $\alpha_i \sim \mathcal{U}(0.8, 1.2)$. See Table 1 for the assumed true medians (before applying the study-specific random effects) for each of the 12 unique combinations of covariates. We then simulated event times for $N_i = 20$ subjects for each of the 50 cohorts, using the distributions indicated in Table 1. Finally, using the `survfit` function in the *R* package *survival* (Therneau, 2023), we evaluated point estimates m_i and corresponding 95% confidence intervals (ℓ_i, h_i) for median PFS for each cohort. The triples $\boldsymbol{s}_i$ are the data. For simplicity, we omitted censoring.

We used the same prior distributions as for the data analysis with the real data, namely $c = 5$ and a half Cauchy with scale 3.5 for G_0. We construct a correlation function $R(\boldsymbol{x}, \boldsymbol{x}')$ as in Section 5, using only the rules for tumor type, agent, and biomarker status. We add future cohorts, $i = n + 1, \ldots, I$, including one cohort for each of the 12 unique combinations of covariates, ie, $I = n + 12$. The choice of $R(\boldsymbol{x}, \boldsymbol{x}')$ is an important step in the model construction, representing informative prior expert judgment. To explore the impact of this choice, we carried out some sensitivity analysis with alternative correlation functions on the same simulated data. These simulations are summarized in the online Supplementary Materials. The simulations show noticable differences in inference under alternative constructions, confirming the importance of an expert-informed construction, as described. To summarize posterior inference on specific tumor-agent pairs, we evaluate log ratios of median PFS, $\log(M_i^+/M_i^-)$, for the 12 future cohorts, comparing marker-positive and -negative pairs (i^+, i^-) of cohorts with matching tumor type and agent. We also evaluate an overall effect of biomarker status by considering a mixture of the 6 marker-positive (P^+) and the 6 marker-negative (P^-) future cohorts using weights proportional to S_i in Table 1. We evaluate point estimates as posterior medians of $\log(M_i^+/M_i^-)$ and $\log(M_P^+/M_P^-)$, respectively. We used log median ratios because those are used in the classical meta-regressions that we report for comparison. For the latter, we used a random effects model with just the intercept (ie, that estimates the overall effect) and a meta-regression including an interaction of tumor type and agent (ie, including a parameter for each the six combinations). The data are the log ratios of median survival times reported in the studies (Michiels et al., 2005). For both models, we used the implementation in the `R` package `metafor` (Viechtbauer, 2010).

Figure 2 shows box-plots of the bias for the different estimates for the log median ratios. Each boxplot shows the realized bias of the posterior estimated log ratios over 50 repeat simulations. For comparison, we also report estimates under a classical meta-regression. The classical meta-regression results show high variability in the estimates across simulations, which is substantially reduced by the proposed model-based inference. This is due to shrinkage towards the prior and the borrowing of strength across similar cohorts. The increased precision is achieved despite the greater flexibility and in the absence of parametric assumptions in the multivariate mvPT model. Figure 2 suggests that this is due to a trade-off with increased average bias over simulations under

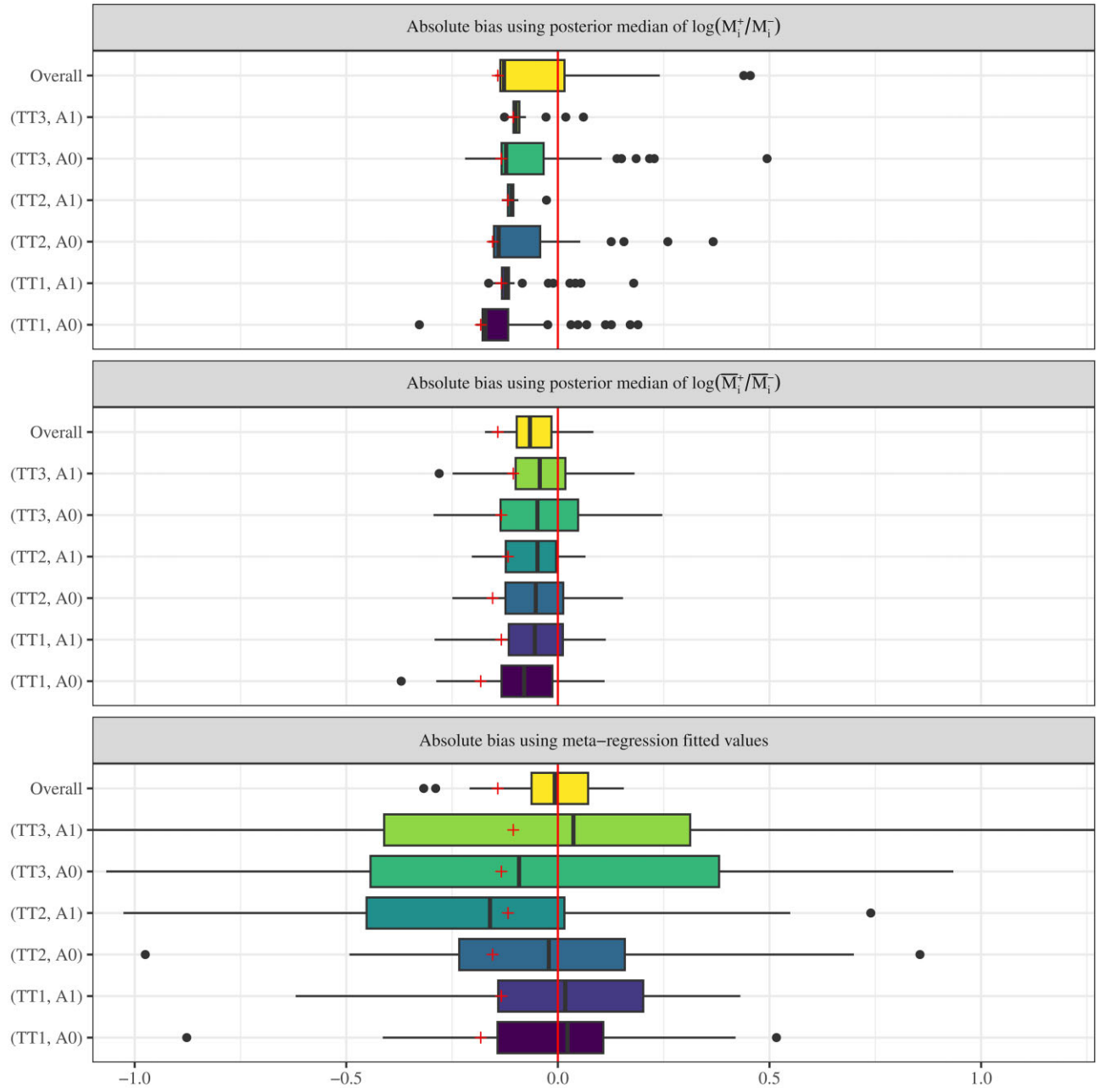

FIGURE 2 Box-plots of realized bias over 50 repeat simulations, for different point estimates. The top panel shows bias for the posterior median of $\log\left(M_i^+/M_i^-\right)$. The middle panel shows bias for the posterior median of $\log(\bar{M}_i{}^+/\bar{M}_i{}^-)$. The bottom panel shows bias for log median ratios under meta-regression approach. For meta-regression bias is calculated using the fitted expected values. In all three plots, the + marks the bias of prior expected values (ie, $\log(1) = 0$). The shifts are due to the different simulation truths of the different covariate combinations.

the proposed approach. More summaries of absolute bias across simulations are available in the online Supplementary Materials. We avoid comparing results between tumor-agent pairs, as the interpretation of such comparisons hinge on the specific simulation truth and sample sizes, and thus would require substantially more simulations scenarios. Finally, keeping in mind that the main inference target is the evaluation of the hypothesis $M_i^+ > M_i^-$, we also assessed methods in terms of inference on this hypothesis by comparing Bayes factors and Bayes factor bounds (Sellke et al., 2001) for classical inference. The latter are evaluated using p-values from a meta-regression. These results are shown in the online Supplementary Materials.

7 RESULTS FOR THE CANCER IMMUNOTHERAPY META-ANALYSIS

Overall, the results support a recommendation for including relevant biomarkers in the design of cancer immunotherapy studies. For almost all tumor-agent pairs, we find high posterior

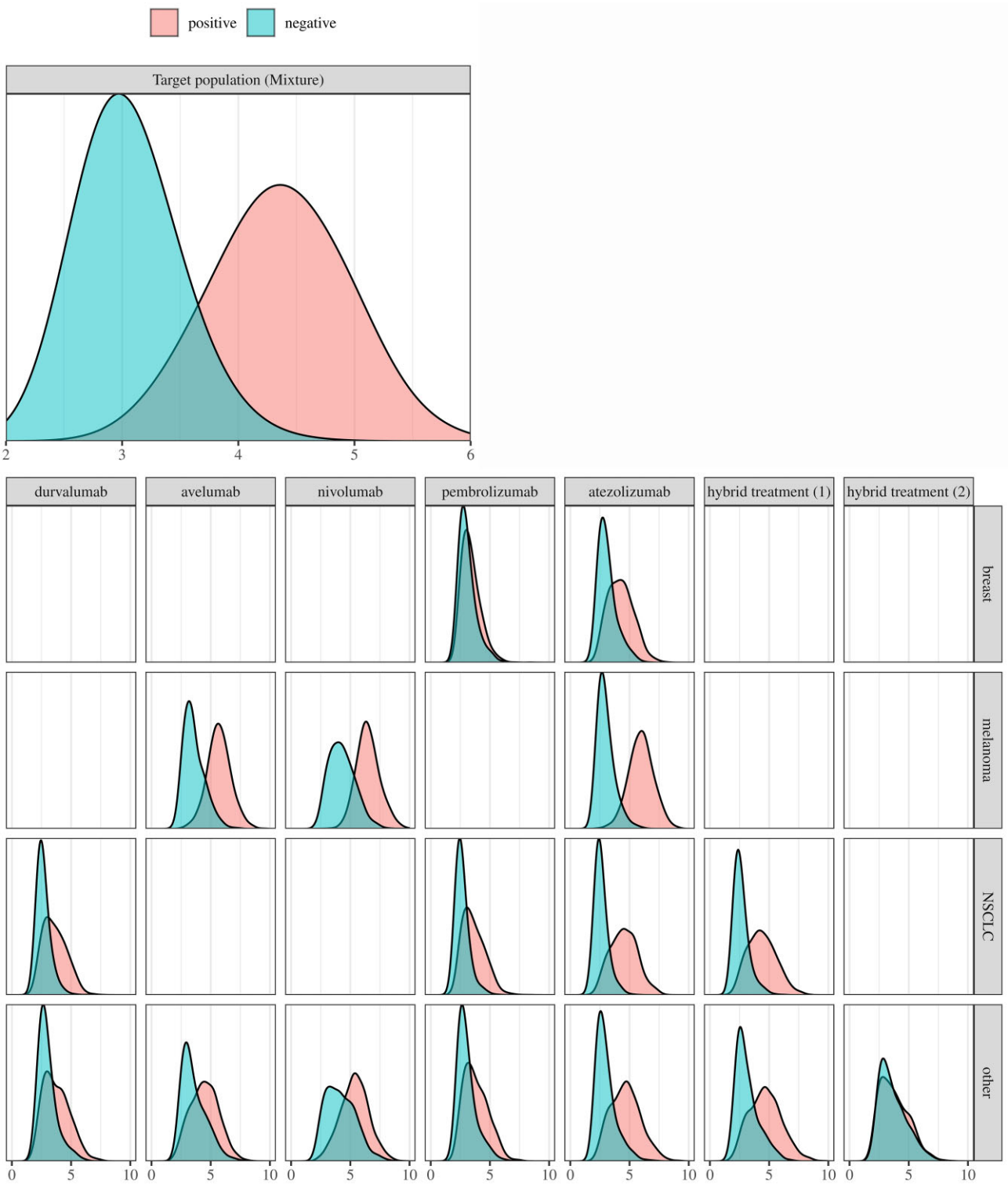

FIGURE 3 Immunotherapies. The top panel shows posterior distributions for M_P^+ and M_P^- for a hypothetical future study, averaging over study characteristics. The lower (small) panels show the same for M_i^+ and M_i^- for hypothetical future studies, arranged by agent and tumor type. Here *hybrid treatment* refers to studies with multiple agents, with (1) for ipilimumab or nivolumab and (2) for pembrolizumab or nivolumab. Empty facets indicate there was no study for the tumor agent pair.

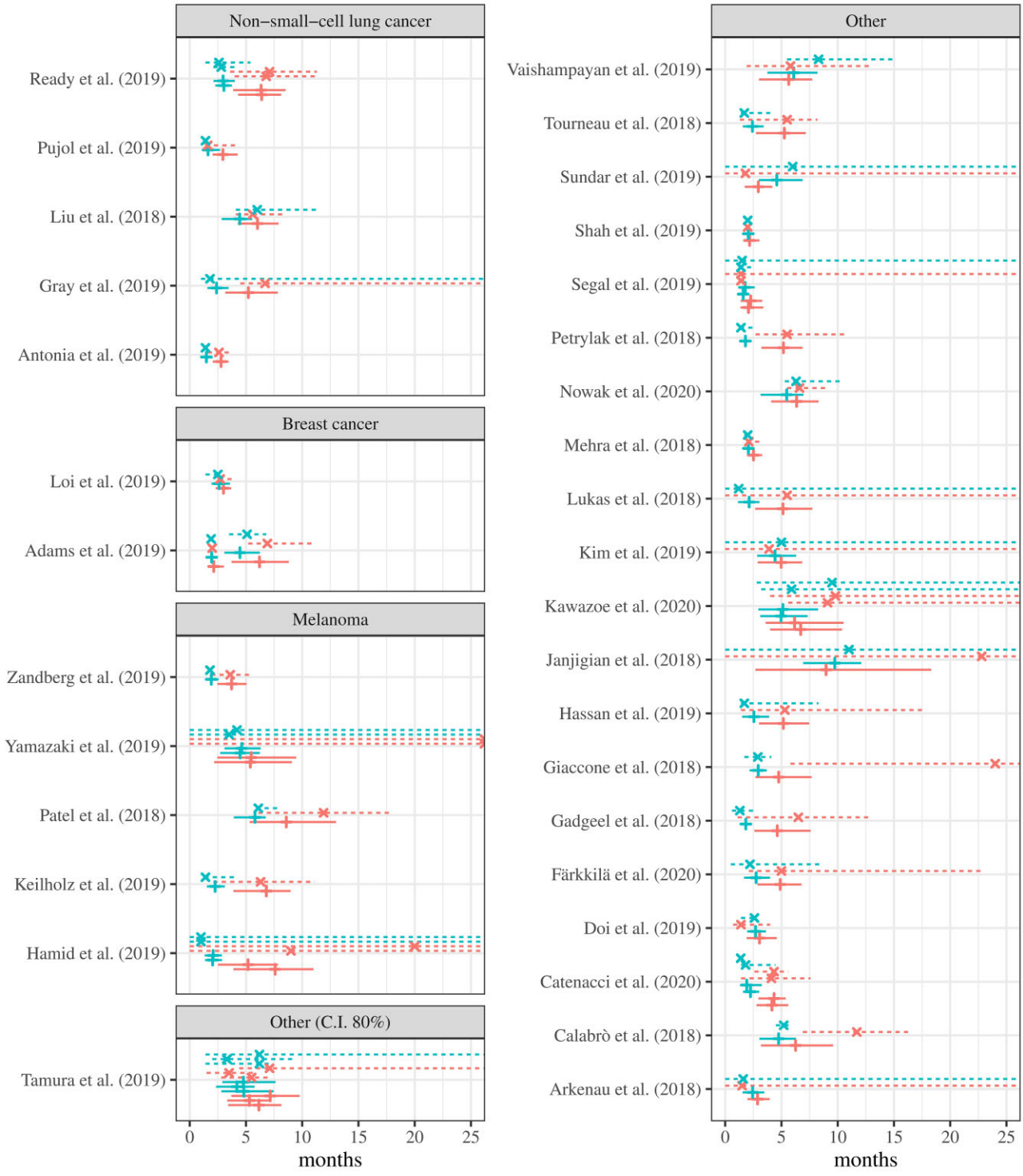

FIGURE 4 Immunotherapies. Summary of posterior inference for the included studies. Intervals represented with + and solid lines are credible intervals for median PFS. Intervals represented with × and dotted lines are confidence intervals reported in the original articles. Only for original summaries, non-observed lower limits are marked with 0 and non-observed upper limits with $+\infty$. The confidence level for all intervals is 95% unless otherwise specified. Marker-positive and marker-negative cohorts are marked in different colours. .

probability for higher median PFS for marker-positive than for marker-negative patients, ie, for $M_P^+ > M_P^-$. Here, $P = \sum \pi_i G_i$, for a mixture over all future studies with maker-positive and -negative status, respectively, as described in Section 4.3. Figure 3 reports inference for hypothetical future studies, arranged by combinations of tumor and agent (small panels) and overall (large panel). Omitted minor covariates (except for biomarker status) are fixed at the most common observed level. Web Figure 6 in the supporting materials reports the same as Figure 3, but for $\overline{M}_P^+$ and $\overline{M}_P^-$.

Our meta-analysis of PFS in immunotherapy studies supports the hypothesis that biomarkers can be useful to identify patients who will benefit from immune checkpoint inhibitors. However, both the effect size and confidence in the results vary depending $\boldsymbol{x}_i$. In particular, there is evidence for an interaction of agent and tumor type. The results are strongest for melanoma and robust for non-small cell lung cancer (NSCLC). For breast cancer (only two studies) and cancers classified as "other" results are concordant with clinical practice but with less strong evidence. We find a protective effect for most agents. The results for atezolizumab and avelumab are the most consistent across tumors. In contrast, for ipilimumab or nivolumab (*hybrid treatment*) and pembrolizumab, the data does not provide equally strong evidence. Posterior credible intervals (one-sided) are available in the online Supplementary Materials. Results on pembrolizumab and breast cancer conflict with clinical experience that would suggest an effect. This is likely due to the sample sizes of the cohorts in the study reported by Adams et al. (2019). These cohorts report equally low estimates for both medians, M_i^+ and M_i^-, with large sample sizes (105 marker-positive and 67 marker-negative patients), which implies higher shrinkage of the medians of future cohorts with breast cancer and pembrolizumab towards these point estimates. This observation highlights the importance of the covariance function and its local behavior.

Figure 4 summarizes inference for the original studies included in Fountzilas et al. (2023b). For each study, the figure compares posterior credible intervals for M_i^+ and M_i^- together with the confidence intervals in the original papers. For most studies, the posterior credible intervals for M_i^+ are higher than those for M_i^-. In particular, for some studies that reported $M_i^+ < M_i^-$ in the original papers, posterior inference switched the order of the point estimates borrowing strength across all cohorts (Liu et al., 2018; Tamura et al., 2019; Segal et al., 2019; Kim et al., 2019; Doi et al., 2019, and Arkenau et al., 2018 in Fig. 4). The reverse only occurs in one case, for Janjigian et al. (2018), which is a study reporting extreme values for the medians, under unusually high and unequal sample sizes (100 marker-positive and 289 marker-negative patients, respectively).

8 CONCLUSIONS

We developed a nonparametric Bayesian approach for meta-analysis with event-time outcomes. Inference combines information from all studies in the meta-analysis. The approach uses weakly informative priors based on clinical expert judgment regarding the relationship between different studies. A simulation study shows that for realistic sample sizes and data structure inference under the proposed approach compares favorably with standard methods. This is especially true for tumor types and agents that are less commonly observed in the original studies. This is achieved by borrowing strength across all studies.

One limitation of the proposed nonparametric approach is that it is restricted to event-time outcomes. A possible future development would be the inclusion of multiple endpoints. Many of the studies in our data include in addition to PFS also summaries for objective response rate (OR), and overall survival (OS). Sharing information on multiple event time endpoints, like PFS and OS, is easily accommodated by treating them as separate, but highly correlated, cohorts. The inclusion of binary endpoints like OR in a joint model would require model extensions to include a parametric submodel.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Pseudo code, Web Tables and Figures referenced in Sections 3, 6, and 7, and data and code to reproduce the results of the Sections 6 and 7 are available with this paper at the Biometrics website on Oxford Academic.

FUNDING

First author was supported in part by the "Dipartimenti Eccellenti 2023-2027" ministerial funds (Italy). Last author was supported in part by National Science Foundation grants DMS 1952679.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The data, code, and the simulation results that support the findings in this paper are also available on GitHub at link https://github.com/GiovanniPoli/mvPTgp.

REFERENCES

Adams, S., Diamond, J. R., Hamilton, E., Pohlmann, P. R., Tolaney, S. M., Chang, C. W., et al. (2019). Atezolizumab plus nab-paclitaxel in the treatment of metastatic triple-negative breast cancer with 2-year survival follow-up: a phase 1b clinical trial. *JAMA Oncol*, 5, 334–342.

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distributions: some properties and uses. *Biometrika*, 67, 261–272.

Arkenau, H. T., Martin-Liberal, J., Calvo, E., Penel, N., Krebs, M. G., Herbst, R. S., et al. (2018). Ramucirumab plus pembrolizumab in patients with previously treated advanced or metastatic biliary tract cancer: Nonrandomized, open-label, phase I trial (JVDF). *Oncologist*, 23, 1407–e136.

Branscum, A. J. and Hanson, T. E. (2008). Bayesian nonparametric meta-analysis using Pólya tree mixture models. *Biometrics*, 64, 825–833.

Christensen, J. and Ma, L. (2019). A Bayesian hierarchical model for related densities by using Pólya trees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82, 127–153.

Diana, A., Matechou, E., Griffin, J., Arnold, T., Tenan, S. and Volponi, S. (2023). A general modeling framework for open wildlife populations based on the Polya tree prior. *Biometrics*, 79, 2171–2183.

Doi, T., Iwasa, S., Muro, K., Satoh, T., Hironaka, S., Esaki, T., et al. (2019). Phase 1 trial of avelumab (anti-PD-L1) in Japanese patients with advanced solid tumors, including dose expansion in patients with gastric or gastroesophageal junction cancer: the JAVELIN Solid Tumor JPN trial. *Gastric Cancer*, 22, 817–827.

Flores, B. and Müller, P. (2024). Clustering and meta-analysis using a mixture of dependent linear tail-free priors. preprint arXiv:2406.15912. https://doi.org/10.48550/arXiv.2406.15912

Fountzilas, E., Hiep Vo, H., Mueller, P., Kurzrock, R. and Tsimberidou, A.-M. (2023a). Dataset of phase I and II immunotherapy clinical trials

used for a meta-analysis to assess the role of biomarkers in treatment outcomes in diverse cancers. *Data in Brief*, 51, 109698.
Fountzilas, E., Vo, H. H., Müller, P., Kurzrock, R. and Tsimberidou, A.-M. (2023b). Correlation between biomarkers and treatment outcomes in diverse cancers: a systematic review and meta-analysis of phase I and II immunotherapy clinical trials. *European Journal of Cancer*, 189, 112927.
Greenwood, M. (1926). A report on the natural duration of cancer. *Reports on Public Health and Medical Subjects*, London.: H.M.S.O., (iv + 26 pp.).
Hosmer, D. W. Jr, Lemeshow, S. and May, S. (2011). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Hoboken, NJ: Wiley-Interscience.
Janjigian, Y. Y., Bendell, J., Calvo, E., Kim, J. W., Ascierto, P. A., Sharma, P., et al. (2018). Checkmate-032 study: Efficacy and safety of nivolumab and nivolumab plus ipilimumab in patients with metastatic esophagogastric cancer. *J Clin Oncol*, 36, 2836–2844. https://doi.org/10.1200/JCO.2017.76.6212
Jara, A. and Hanson, T. E. (2011). A class of mixtures of dependent tail-free processes. *Biometrika*, 98, 553–566.
Jara, A., Hanson, T. E. and Lesaffre, E. (2009). Robustifying generalized linear mixed models using a new class of mixtures of multivariate Pólya trees. *Journal of Computational and Graphical Statistics*, 18, 838–860.
Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.
Kim, Y. J., Keam, B., Ock, C. Y., Song, S., Kim, M., Kim, S. H., et al. (2019). A phase II study of pembrolizumab and paclitaxel in patients with relapsed or refractory small-cell lung cancer. *Lung Cancer*, 136, 122–128. https://doi.org/10.1016/j.lungcan.2019.08.031 31494530
Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, 20,1222–1235.
Lavine, M. (1994). More aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, 22, 1161–1176.
LeBlanc, P. and Ma, L. (2022). Microbiome subcommunity learning with logistic-tree normal latent Dirichlet allocation. *Biometrics*, 79, 2321–2332.
Liu, S. V., Camidge, D. R., Gettinger, S. N., Giaccone, G., Heist, R. S., Hodi, F. S., et al. (2018). Long-term survival follow-up of atezolizumab in combination with platinum-based doublet chemotherapy in patients with advanced non-small-cell lung cancer. *Eur J Cancer*, 101, 114–122. https://doi.org/10.1016/j.ejca.2018.06.033
Marabelle, A., Le, D. T., Ascierto, P. A., Di Giacomo, A., De Jesus-Acosta, A., Delord, J. P. et al. (2020). Efficacy of pembrolizumab in patients with noncolorectal high microsatellite instability/mismatch repair–deficient cancer: results from the phase II KEYNOTE-158 study. *Journal of Clinical Oncology*, 38, 1.
Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22, 1167–1180.
Michiels, S., Piedbois, P., Burdett, S., Syz, N., Stewart, L. and Pignon, J.-P. (2005). Meta-analysis when only the median survival times are known: a comparison with individual patient data results. *International Journal of Technology Assessment in Health Care*, 21, 119–125.
Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*, Cambrige, MA: MIT Press.
Parmar, M. K., Torri, V. and Stewart, L. (1998). Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine*, 17, 2815–2834.
Patel, S. P. and Kurzrock, R. (2015). PD-L1 expression as a predictive biomarker in cancer immunotherapy. *Molecular Cancer Therapeutics*, 14, 847–856.
Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108, 1339–1349.
Rasmussen, C. E. and Williams, C. K. (2006) *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series)*, Cambridge, MA: MIT Press.
Ruberu, T. L. M., Braun, D., Parmigiani, G. and Biswas, S. (2024). Meta-analysis of breast cancer risk for individuals with PALB2 pathogenic variants. *Genet Epidemiol*. https://doi.org/10.1002/gepi.22561
Schwarzer, G., Carpenter, J. R. and Rücker, G. (2015) *Meta-Analysis with R*. Cham, Switzerland: Springer International Publishing, 21–53.
Segal, N. H., Ou, S.I., Balmanoukian, A., Fury, M. G., Massarelli, E., Brahmer, J. R., et al. (2019). Safety and efficacy of durvalumab in patients with head and neck squamous cell carcinoma: results from a phase I/II expansion cohort. *Eur J Cancer* 109, 154–161. https://doi.org/10.1016/j.ejca.2018.12.029
Sellke, T., Bayarri, M. and Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, 55, 62–71.
Sutton, A. J. and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10, 277–303.
Tamura, K., Hasegawa, K., Katsumata, N., Matsumoto, K., Mukai, H., Takahashi, S., et al. (2019). Efficacy and safety of nivolumab in Japanese patients with uterine cervical cancer, uterine corpus cancer, or soft tissue sarcoma: Multicenter, open-label phase 2 trial. *Cancer Sci*, 110, 2894–2904. https://doi.org/10.1111/cas.14148
Therneau, T. M. (2023). *A Package for Survival Analysis in R. R package version 3.5-7*. https://doi.org/10.32614/CRAN.package.survival
Trippa, L., Müller, P. and Johnson, W. (2011). The multivariate beta process and an extension of the Pólya tree model. *Biometrika*, 98, 17–34.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor Package. *Journal of Statistical Software*, 36, 1–48.
Wang, Z., Mao, J. and Ma, L. (2022). Microbiome compositional analysis with logistic-tree normal models. preprint, https://arxiv.org/abs/2106.15051v4.
Yang, M., Hanson, T. and Christensen, R. (2008). Nonparametric Bayesian estimation of a bivariate density with interval censored data. *Computational Statistics and Data Analysis*, 52, 5202–5214.

APPENDIX A: POSTERIOR UPDATE VIA PÒLYA GAMMA MODEL AUGMENTATION

Let $N_{\varepsilon_d}^{(i)}$ denote the number of observations in the interval $B_{\varepsilon_d}^{(i)}$. The posterior distribution for $\mathbf{Z}_{\varepsilon_d 0} = \{Z_{\varepsilon_d 0}^{(i)}\}_{i=1}^{I}$ depends on $N_{\varepsilon_d}^{(i)}$ and $N_{\varepsilon_d 0}^{(i)}$. The conditional posterior distribution closely resembles a logistic regression. Let A_{ε_d} denote the set of all cohorts i for which at least one observation is recorded in $B_{\varepsilon_d}^{(i)}$, and split $\mathbf{Z}_{\varepsilon_d 0}$ according to A_{ε_d} into $\mathbf{Z}_{\varepsilon_d 0} = \mathbf{Z}_{\varepsilon_d 0}^{1} \cup \mathbf{Z}_{\varepsilon_d 0}^{\varnothing}$. Noting that only studies $i \in A$ contribute likelihood factors, this implies

$$p(\mathbf{Z}_{\varepsilon_d 0} \mid \cdot) \propto \left[\prod_{i \in A_{\varepsilon_d}} \frac{\left(\exp\left\{ Z_{\varepsilon_d 0}^{(i)} \right\} \right)^{N_{\varepsilon_d 0}^{(i)}}}{\left(1 + \exp\left\{ Z_{\varepsilon_d 0}^{(i)} \right\} \right)^{N_{\varepsilon_d}^{(i)}}} \right] p(\mathbf{Z}_{\varepsilon_d 0}^{1})\, p(\mathbf{Z}_{\varepsilon_d 0}^{\varnothing} \mid \mathbf{Z}_{\varepsilon_d 0}^{1}).$$

The data augmentation strategy of Polson et al. (2013) is then implemented as follows:

(1) Sample ω_i from $p(\omega_i \mid \mathbf{Z}_{\varepsilon_d \mathbf{0}}^{\mathbf{1}})$ for each $i \in A_{\varepsilon_d}$ independently using a Pòlya gamma.
(2) Update $Z_{\varepsilon_d 0}^{(i)}$ for each $i \in A_{\varepsilon_d}$ sampling from $p(\mathbf{Z}_{\varepsilon_d 0}^{1} \mid \boldsymbol{\omega})$.

(3) Update $Z^{(i)}_{\varepsilon_d 0}$ for each $i \notin A_{\varepsilon_d}$ by sampling from $p(\boldsymbol{Z}^{\varnothing}_{\varepsilon_d 0} \mid \boldsymbol{Z}^{1}_{\varepsilon_d 0})$. The latter takes the form of a conditional multivariate normal.

APPENDIX B: POSTERIOR UPDATE OF (LATENT) COUNTS

Consider a generic cohort i with (known) sample size N. Omitting the study-specific indices i and (i), let $\boldsymbol{C}_{1:N}$ and $\boldsymbol{T}_{1:N}$ denote the unknown patient level censoring and event times. We first consider the conditional distribution of the complete data, assuming that both censoring and event-time distribution are known. Let $H(C_j)$ denote the earlier, and $G(T_j)$ the latter, and let $\boldsymbol{s}$ denote the triple (ℓ, m, h) implied by $\boldsymbol{T}_{1:N}$, $\boldsymbol{C}_{1:N}$ and $\boldsymbol{s}^o$ the observed data. Then conditioning on all currently imputed parameters, including in particular G itself, and the observed data we have

$$p\left(\boldsymbol{T}_{1:N}, \boldsymbol{C}_{1:N} \mid \cdot\right) \propto p\left(\boldsymbol{s}^o \mid \boldsymbol{T}_{1:N}, \boldsymbol{C}_{1:N}\right) \prod_{j=1}^{N} G\left(T_j\right) H(C_j)$$

$$= \mathbb{I}(\boldsymbol{s} = \boldsymbol{s}^o) \prod_{j=1}^{N} G(T_j)\, H(C_j).$$

We construct Markov chain Monte Carlos simulation to impute $(\boldsymbol{C_{1:N}}, \boldsymbol{T_{1:N}})$ using two transition probabilities defined with the following proposal distributions. Proposal Q_1 proposes a new event and censoring time for two randomly selected subjects r and s in the cohort. We use T and C without superscript for the proposal, and T^o, C^o for the currently imputed values. $Q_1(\boldsymbol{T}_{1:N}, \boldsymbol{C}_{1:N} \mid \boldsymbol{T}^o_{1:N}, \boldsymbol{C}^o_{1:N}) = H\big(C_s\big)G\big(T_s\big) \cdot H\big(C_r\big)G\big(T_r\big)/(n(n-1))$ implying acceptance probability

$$\alpha = \min\left\{1, \frac{\mathbb{I}(\boldsymbol{s} = \boldsymbol{s}') \prod_{j=1}^{N} G\left(T_j\right) H(C_j)}{\prod_{j=1}^{N} G\left(T^o_j\right) H(C^o_j)} \frac{H\big(C^o_s\big)G\big(T^o_s\big) \times H\big(C^o_r\big)G\big(T^o_r\big)}{H\big(C_s\big)G\big(T_s\big) \times H\big(C_r\big)G\big(T_r\big)}\right\} = \mathbb{I}(\boldsymbol{s} = \boldsymbol{s}^o).$$

Proposal Q_2 randomly selects a subject s and proposes to "*flip*" the censoring indicator $\delta_s = \mathbb{I}(T_s < C_s)$ and thereby changing the censoring pattern $\boldsymbol{\delta}$. Let $\tilde{T}_s = \min\{T_s, C_s\}$ denote the implied observed times, and let $\tilde{T}_{(s-1)}$ and $\tilde{T}_{(s+1)}$ denote the largest observed time less than $\tilde{T}_s$ and the smallest one greater than $\tilde{T}_s$, respectively (all quantities are known by conditioning on the currently current imputed values). Then

$$\begin{aligned} Q_2(\boldsymbol{T}_{1:N}, \boldsymbol{C}_{1:N} \mid \boldsymbol{T}^o_{1:N}, \boldsymbol{C}^o_{1:N}) &= \frac{1}{N}\left[H\left(C_s \mid \tilde{T}^o_{(s-1)} < C_s < \tilde{T}^o_{(s+1)}\right) G\left(T_s \mid C_s < T_s, C_s\right)\right]^{\delta^o_s} \\ &\quad \times \left[G\left(T_s \mid \tilde{T}^o_{(s-1)} < T_s < \tilde{T}^o_{(s+1)}\right) H\left(C_s \mid T_s < C_s, T_s\right)\right]^{1-\delta^o_s} \\ &= \frac{1}{N}\left[\frac{H\left(C_s\right)}{H\left(\tilde{T}^o_{(s-1)} < C_s < \tilde{T}^o_{(s+1)}\right)} \frac{G\left(T_s\right)}{G\left(T_s > C_s \mid C_s\right)}\right]^{\delta^o_s} \\ &\quad \times \left[\frac{G\left(T_s\right)}{G\left(\tilde{T}^o_{(s-1)} < T_s < \tilde{T}^o_{(s+1)}\right)} \frac{H\left(C_s\right)}{H\left(C_s > T_s \mid T_s\right)}\right]^{1-\delta^o_s}. \end{aligned} \tag{B1}$$

This implies the acceptance probability $\alpha_2 = \min(1, r)$ with

$$r = \mathbb{I}\left(\boldsymbol{s} = \boldsymbol{s}^o\right) \begin{cases} \dfrac{G\left(\tilde{T}^o_{(s-1)} < T_s < \tilde{T}^o_{(s+1)}\right) H\left(C_s > T_s \mid T^{new}_s\right)}{H\left(\tilde{T}^o_{(s-1)} < C_s < \tilde{T}^o_{(s+1)}\right) G\left(T_s > C_s \mid C^o_s\right)} & \text{if } \delta^o_s = 1 \\ \dfrac{H\left(\tilde{T}^o_{(s-1)} < C_s < \tilde{T}^o_{(s+1)}\right) G\left(T_s > C_s \mid C^{new}_s\right)}{G\left(\tilde{T}^o_{(s-1)} < T_s < \tilde{T}^o_{(s+1)}\right) H\left(C_s > T_s \mid T^o_s\right)} & \text{if } \delta^o_s = 0. \end{cases}$$

We divided the cohorts into three groups. (1) Cohorts with known censoring pattern (ie, no censoring or censoring only after the last event time); (2) cohorts with reported number of events or number of censored outcomes; (3) cohorts for which only the sample size is recorded. For (1) counts are not updated since they are known. For (2) the filp step is never pro-

posed. For (3) one of the two proposals is randomly selected each time.

To initialize $\boldsymbol{T}_{1:N}$ and $\boldsymbol{C}_{1:N}$, we start with a censuring pattern where all event times are observed and we followed the following argument. Let $T_{(j)}$ denote the order statistic for $\boldsymbol{T}_{1:N}$. We can then consider a Kaplan–Meier plot with the index j on the horizontal axis (that is, plotting against the indices of the order statistic instead of the unknown actual times). We plot the estimated survival function and $(1-\alpha)$ confidence interval bands (still plotted against j). Assuming w.l.o.g. $N = 2k+1$, and assuming that the recorded data $\boldsymbol{s} = (\ell, m, h)$ are determined as the intersections of the three curves with the 0.5 threshold, we can then identify the data (ℓ, m, h) as $(T_{(L)}, T_{(k+1)}, T_{(H)})$. The remaining data T_j are generated from G, subject to the given $(T_{(L)}, T_{(k+1)}, T_{(H)})$. That is, we generate $L-1$ event times $T_j \sim G \cdot \mathbb{I}(T_j < T_{(L)})$ etc.

APPENDIX C: CORRELATION MATRIX

Recall the construction of $R(\boldsymbol{x}, \boldsymbol{x}')$ in §5. Consider two cohort-specific covariate vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$, and denote with $k_1(\boldsymbol{x}, \boldsymbol{x}')$ the similarity score based on matching covariates.

- $k_1(\boldsymbol{x}, \boldsymbol{x}')$ be a linear kernel with coherent weights obtained using the one-hot encoding of the categorical variables.
- $k_2(\boldsymbol{x}, \boldsymbol{x}')$ be an indicator for $\boldsymbol{x}$ and $\boldsymbol{x}'$ having matching marker status.
- $k_3(\boldsymbol{x}, \boldsymbol{x}')$ be an indicator for $\boldsymbol{x}$ and $\boldsymbol{x}'$ sharing the same study.
- $k_4(\boldsymbol{x}, \boldsymbol{x}')$ be the identity kernel, ie, an indicator for $\boldsymbol{x}$ and $\boldsymbol{x}'$ referring to the same cohort.

We combine the kernels to obtain the covariance functions described in §5:

$$R(\boldsymbol{x}, \boldsymbol{x}') = \frac{k_1(\boldsymbol{x}, \boldsymbol{x}') \cdot k_2(\boldsymbol{x}, \boldsymbol{x}') + k_3(\boldsymbol{x}, \boldsymbol{x}') + k_4(\boldsymbol{x}, \boldsymbol{x}')}{8}.$$

$\boldsymbol{R}$ is positive definite since the sum and the product of positive semi-definitive kernels are positive semi-definite and $k_4(\boldsymbol{x}, \boldsymbol{x}')$ is positive definite.

For other choices of building the similarity score, the construction $R(\boldsymbol{x}, \boldsymbol{x}')$ might not be p.d. In that case, it is always possible to scale $k_4(\boldsymbol{x}, \boldsymbol{x}')$ by some $c > 0$ to ensure R to be p.d. This is the case due to the following result. If A is a symmetric $(I \times I)$ matrix, then there is a $c > 0$ such that $B = [A + c \cdot I_{I \times I}]$ is p.d. The result is easy to prove by considering the normalized eigenvalues of B. See also Rasmussen and Williams (2006) for a discussion of covariance functions, keeping in mind that the requirement for p.d. covariance matrices is only needed for the I cohorts under consideration.

Received: December 12, 2023; Revised: October 3, 2024; Accepted: October 29, 2024