

TITLE: Joint Analysis of Two Time-to-Event Outcomes Using a Product Partition Model Accounting for Semi-competing Risks

AUTHOR(S): Kang Lei, Satoshi Morita, Shogo Nakamoto, Junichiro Watanabe

Version of Record online: 05 April 2025

Copyright and Acknowledgement: This is the peer reviewed version of the following article: Kang Lei, Satoshi Morita, Shogo Nakamoto & Junichiro Watanabe. Joint Analysis of Two Time-to-Event Outcomes Using a Product Partition Model Accounting for Semi-competing Risks.

This version of the article has been accepted for publication, after peer review (when applicable) but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: <https://doi.org/10.1007/s12561-025-09480-x>. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use <https://www.springernature.com/gp/open-science/policies/accepted-manuscript-terms>.

Joint Analysis of Two Time-to-Event Outcomes using a Product Partition Model Accounting for Semi-Competing Risks

Kang Lei^{1*}, Satoshi Morita¹, Shogo Nakamoto²,
Junichiro Watanabe³

¹Department of Biomedical Statistics and Bioinformatics, Graduate
School of Medicine, Kyoto University, Kyoto, Japan.

²Department of General Thoracic Surgery and Breast and
Endocrinological Surgery, Graduate School of Medicine, Okayama
University, Okayama, Japan.

³Department of Breast Oncology, Graduate School of Medicine,
Juntendo University, Tokyo, Japan.

*Corresponding author(s). E-mail(s): kanglei_adam@outlook.com;

Abstract

Cancer clinical trials often encompass various time-to-event (TTE) outcomes, such as overall survival, progression-free survival, and time-to-treatment failure, with each shedding light on related but different aspects of treatment efficacy. However, given the intricate relationships and mechanisms by which survival is extended for patients with cancer, making suitable treatment decisions for patients becomes challenging. To address this challenge and enhance precision in treatment selection, identifying patient subgroups before commencing treatment is important. To achieve this objective, concurrently considering two TTE outcomes and employing clustering analysis to determine the patient subgroups, thereby facilitating the selection of the most appropriate treatment, might be a more suitable approach. In this paper, on the basis of the random partition model with regression on covariates (PPMx), we propose a PPMx based clustering model (DbTTE-PPMx) with Gaussian copula to jointly model two TTE outcomes while accounting for their correlations. Importantly, we account for the presence of semi-competing risks to consider the complicated interplay between these two TTE measures. In addition, we examine the operating characteristics

of the proposed method in simulation studies. For illustration, we apply the proposed method to data from a clinical study in patients with advanced breast cancer.

Keywords: Product partition model, two time-to-event outcomes, semi-competing risks, Gaussian copula, cancer clinical study

1 Introduction

In cancer clinical research, Pazdur [1] explained that various endpoints, such as overall survival (OS), progression-free survival (PFS), and time-to-treatment failure (TTF) are critical outcomes to consider. Prolonging the time that a patient remains free from disease progression is particularly important in the context of cancer care, because it enables individuals to lead a higher quality of life. Simultaneously, the possibility of death is also a significant concern for both the patient and the clinician. However, the relationship between these time-to-event (TTE) outcomes might not be straightforward. Notably, achieving a highly significant result in terms of PFS might not necessarily translate into a similarly significant outcome in terms of OS [2][3] [4]. This observation indicates that the complicated relationship between these clinical endpoints makes the decision-making process more challenging. Moreover, in practice, substantial population-level heterogeneity often exists because of unknown subgroup structures. Consequently, the interpretation of cluster structures is a critical aspect of survival analysis.

Our objective is to identify common characteristics shared among the patient population and determine the subgroup of patients to provide useful information for clinical decision making. We assume that the patient population can be partitioned into different subgroups. Under this assumption, we propose that the product partition framework might be more suitable for interpreting such partitioned subgroup structures [5]. Moreover, to incorporate the information of patients' covariates, we analyze the cluster structure of patients by using the product partition model with regression on covariates (PPMx), as introduced by Müller et al. [6]. This approach enables us to examine how covariates are associated with a TTE outcome by incorporating the dependence on covariates. However, the model is limited by its ability to only accommodate a single TTE outcome at a time. [Previous work on modelling the cluster structure of patients using a product partition framework focused more on modelling covariates. Page et al.\[7\] focused on the product partition model to handle missing covariates, while Pedone et al.\[8\] extended the product partition model framework by considering prognostic and predictive biomarkers.](#)

Traditional approaches often involve a separate analysis of each TTE outcome, which might not provide a complete understanding of how patient outcomes are related to each other. This limitation arises from the observation that the membership assignments are derived from the separate analysis of two distinct TTE outcomes that are independent of each other, making the concurrent interpretation of the membership from two different outcomes using the existing methods impossible. In such a case, we

cannot interpret a meaningful cluster structure from the results. The situation also occurs in our application dataset when conducting inference in a clinical study, in which two TTE outcomes are investigated after patients receive their anticancer treatment with a number of baseline covariates for each patient. For example, as exemplified in section 4, the traditional approach might well depict a complicated relationship between a covariate and both TTF and OS with different cluster structures separately. This situation can underscore the challenge of interpreting the cluster structures of patients when two TTE outcomes require concurrent consideration. Previous studies on the joint analysis of two TTE outcomes have explored various aspects but have certain limitations. Alternative approaches, such as [9] [10], they focused on binary TTE outcomes but without considering patient covariates. Other approaches [11] [12] modelled the relationship between two TTE outcomes by considering the covariates. However, they have not provided the cluster structures of patients which may reveal common characteristics shared among the patient populations. Motivated by these compelling reasons, instead of simply analyzing the cluster structure using a single TTE outcome, this study extended the PPMx model by using a Gaussian copula to provide a reasonable interpretation by analyzing two TTE outcomes simultaneously.

In addition, when conducting an analysis involving two TTE outcomes, the presence of semi-competing risks might be important to be considered [13]. Unlike scenarios involving standard competing risks, in a semi-competing risks setting, it might be possible to observe both nonterminal and terminal events in the same patients. Xu et al. [14] illustrated that this unique aspect implies the potential existence of some partial information on the joint distribution of these events. To address this complex challenge and utilize most of the available information, our proposed double-TTE-outcomes PPMx (DbTTE-PPMx) method enables the concurrent analysis of both TTE outcomes while accounting for the semi-competing risks problem making it highly valuable in clinical settings. By jointly analyzing both TTF and OS, we may identify the patient subgroups with different patterns of treatment responses. On the basis of the subgroups, we may gain a better understanding of the effect of treatments on, for instance, patient outcomes and subgroup structure. These findings might enable clinicians to make more informed decisions about treatment options for patients. The remainder of this paper is organized as follows. Section 2 describes and presents our proposed DbTTE-PPMx method for jointly modelling two TTE outcomes in the framework of the semi-competing risks. In section 3, we conduct extensive simulation studies to examine the operating characteristics of the proposed method. In section 4, we present an application of the proposed method. The paper closes with a brief discussion in section 5.

2 Methods

In our methodological exploration, we focus on the product partition model (PPM), which was introduced by Hartigan [15], Barry and Hartigan [5] and Quintana and Iglesias [16]. The PPM constructs the prior probability of partition using the cohesion function, which has the advantage of reflecting how tightly the elements are grouped

in the clusters. Additionally, the PPM could provide a representation of data distribution as the product form of different independent clusters, showing its advantage of simplicity when modelling heterogenous data. Subsequently, Müller et al. [6] extended the PPM by considering the regression on covariates; their model is referred to as PPMx. By leveraging these strengths, we aim to extend PPMx to jointly analyze two TTE outcomes. In this section, we will initially introduce the notation and further elucidate how the PPMx is extended to jointly analyze two TTE outcomes and address semi-competing risks challenges. Finally, we will explain how the posterior inference is summarized.

2.1 Notation

Let $i = 1, \dots, n$ be the index of patients and let $\rho_n = \{S_1, \dots, S_{k_n}\}$ denote a partition of the n patients into k_n subsets S_j . Let x_i denote the patient-specific covariates and $\mathbf{X} = (x_1, \dots, x_n)$ be the complete set of covariates. In the case of a single TTE outcome, let Y denote the TTE outcome for the n patients. In the case involving two TTE outcomes, let the two distinct TTE outcomes of patient i be denoted as $y_{S,i}$ and $y_{L,i}$, where $y_{S,i} \leq y_{L,i}$. Let $Y_S = (y_{S,1}, \dots, y_{S,n})$ and $Y_L = (y_{L,1}, \dots, y_{L,n})$ be two sets of TTE outcomes. Let $\mathbf{Y} = \{Y_S, Y_L\}$ represent the entire set for two different TTE outcomes. Let $\{\gamma_{S,i}, \gamma_{L,i}\}$ be the indicator variables for $y_{S,i}$ and $y_{L,i}$, where the value 1 is assumed when the event is observed and the value 0 is taken otherwise. Let $\mathbf{x}_j^* = (x_i, i \in S_j)$ and $\mathbf{y}_j^* = (y_i, i \in S_j)$ denote the covariates and TTE outcomes arranged by clusters.

2.2 Product partition model for two time-to-event outcomes

In the context of the PPM, the probability model for the response data Y can be expressed as

$$p(Y) = \prod_{j=1}^{k_n} p_{S_j}(\mathbf{y}_j^*) \quad (1)$$

and the prior probability which is assigned to partition ρ_n is given by

$$p(\rho_n) \propto \prod_{j=1}^{k_n} c(S_j) \quad (2)$$

where $c(S_j) \geq 0$ denotes the cohesion function, which measures the strength in how tightly the elements are grouped in S_j . One simple form can be defined as depending only on the number of subjects in S_j , which favor the large size of the subset. A commonly used form, derived from the Dirichlet Process (DP), is $c(S_j) = M(|S_j| - 1)$ where M denotes the mass parameter of DP. The PPM is then specified through equations (1) and (2). It is noted that the corresponding posterior distribution of ρ is obtained in a same product form of (2). The PPM is extended to the PPMx by introducing an additional factor $g(\mathbf{x}_j^*)$ to equation (2) to achieve the dependence on

the covariates [6]. It can be defined as

$$p(\rho_n | \mathbf{X}) \propto \prod_{j=1}^{k_n} g(\mathbf{x}_j^*) \cdot c(S_j) \quad (3)$$

where $g(\cdot)$ denotes a non-negative function of $\mathbf{x}_j^*$ that formalizes the similarity of covariates among patients, with higher values $g(\cdot)$ assigned to sets of covariates that exhibit greater similarity. As a default choice from [6], $g(\cdot)$ is the marginal probability in an auxiliary probability model $q(\cdot)$, which can be represented as

$$g(\mathbf{x}_j^*) \equiv \int \prod_{i \in S_j} q(x_i | \xi_j) q(\xi_j) d\xi_j$$

where $q(\cdot)$ is defined as the probability density that depends on the type of covariates. The correlation among covariates is induced by the cluster-specific parameters ξ_j , leading to higher values of $g(\mathbf{x}_j^*)$ for tightly clustered covariates. For example, for continuous covariates, the marginal distribution of $\mathbf{x}_j^*$ can be defined as a normal sampling model. Let $N(x|m, V)$ be the normal model for the random variable x , with mean m and variance V and let $Ga(x|a, b)$ be a gamma model for random variable x with mean a/b and variance a/b^2 . Therefore, the model can be shown as $q(\mathbf{x}_j^*|\xi_j) = \prod_{i \in S_j} N(x_i|m_j, v_j)$, with conjugate prior for the cluster-specific parameters $\xi_j = (m_j, v_j)$ as $p(\xi_j) = N(m_j|m, B) \cdot Ga(v_j^{-1}|\vartheta, D_0)$ with predefined fixed values for m, B, ϑ, D_0 . For categorical covariate, a default choice is on the basis of a Dirichlet prior distribution. For count covariates, a default choice is on the basis of a mixture of Poisson distributions. The details are provided in the supplementary materials. When dealing with two TTE outcomes, PPMx is required to analyze two TTE outcomes independently rather than simultaneously. The model is completed by specifying distinct data sampling model for Y_S and Y_L separately as follows:

$$p(Y_S|\rho_n, \boldsymbol{\kappa}, \eta, \mathbf{X}) = \prod_{j=1}^{k_n} \prod_{i \in S_j} p(y_{S,i}|x_i, \kappa_j, \eta)$$

and

$$p(Y_L|\rho_n, \boldsymbol{\kappa}, \eta, \mathbf{X}) = \prod_{j=1}^{k_n} \prod_{i \in S_j} p(y_{L,i}|x_i, \kappa_j, \eta) \quad (4)$$

where $\boldsymbol{\kappa}$ is the cluster specific parameters $\boldsymbol{\kappa} = \{\kappa_1, \dots, \kappa_{k_n}\}$ and common hyper-parameters η . We extend the PPMx to model two TTE outcomes using the Gaussian copula structure which can be shown as $G(Y_S, Y_L|\boldsymbol{\lambda}, \boldsymbol{\kappa}_j) = \psi[\Phi^{-1}\{F_S(y_{S,i}|\lambda_{Sj})\}, \Phi^{-1}\{F_L(y_{L,i}|\lambda_{Lj})\}|\kappa_j]$ where $\boldsymbol{\lambda}_j = \{\lambda_{Sj}, \lambda_{Lj}\}$ denote the parameters in the marginal exponential survival functions for the two TTE outcomes and κ_j is the cluster-specific parameter. Let $\nu_j = \{\kappa_j, \boldsymbol{\lambda}_j\}$, we can obtain the likelihood

for patient i in cluster j :

$$p(\mathbf{Y}|\rho_n, \boldsymbol{\nu}, \mathbf{X}, \eta) = \prod_{j=1}^{k_n} \prod_{i \in S_j} p(y_{S,i}, y_{L,i} | x_i, \nu_j, \eta), \quad p(\boldsymbol{\nu}|\eta) = \prod_{j=1}^{k_n} p(\nu_j|\eta) \quad (5)$$

We complete the model with the combination of model (3) and (5) as DbTTE-PPMx. However, in the context of analyzing two TTE outcomes, the potential challenge of semi-competing risks might arise. Considering and addressing this risk might help us improve understanding of the joint risk of two TTE outcomes, as illustrated by Varadhan et al. [17]. Therefore, this challenge will be well introduced in the next subsection.

2.3 Semi-competing risks modelling

OS time represents the time from treatment initiation to the time of death from any cause. However, longer TTE outcomes such as OS time are not always observable. For instance, patients may drop out of the study early and may not be followed up. For shorter TTE outcomes such as time-to-progression, consider a scenario where disease progression assessments are intermittently missing. A patient is known to be alive without progression at Time 1, and much later dies at Time 2. However, there are no evaluations were conducted between Time 1 and Time 2 to confirm the disease progression status. A terminal event (e.g. death) may preclude the occurrence of a non-terminal event, which is described as a semi-competing risks by [13]. Recall that there are the two TTE outcomes $y_{S,i}$ and $y_{L,i}$, where $y_{S,i} \leq y_{L,i}$. In keeping with the general convention, we consider only the right censoring. The independent, non-informative censoring time for each patient i is denoted as C_i . The four cases of censoring described with event indicators $\{\gamma_{S,i}, \gamma_{L,i}\}$ for patient i are shown below:

1. $y_{S,i}$ and $y_{L,i}$ are both observed: $\gamma_{S,i} = 1, \gamma_{L,i} = 1$
2. $y_{S,i}$ is observed and $y_{L,i}$ is censored: $\gamma_{S,i} = 1, \gamma_{L,i} = 0$
3. $y_{S,i}$ is censored and $y_{L,i}$ is observed: $\gamma_{S,i} = 0, \gamma_{L,i} = 1$
4. $y_{S,i}$ and $y_{L,i}$ are both censored: $\gamma_{S,i} = 0, \gamma_{L,i} = 0$

Suppose that the time from the initiation of chemotherapy treatment/intervention occurs at time 0. To illustrate these cases, we define $A_i = \{y_{S,i} \wedge y_{L,i} \wedge C_i\}$ and $B_i = \{y_{L,i} \wedge C_i\}$, where $\wedge$ is the minimum operator. By considering the relationship between two TTE outcomes in a semi-competing risks framework, we form the likelihood function in four cases on the basis of the approach of Fu et al. [18]. Let $G(\cdot)$ and $g(\cdot)$ be the marginal survival function and the probability density function (pdf) for censoring time C ; $S(y_S, y_L)$ and $f(y_S, y_L)$ are the joint survival function and the joint pdf of $(y_{S,i}, y_{L,i})$, respectively, and the partial derivative expression form of joint survival function can be shown as:

$$\dot{S}_l(y_S, y_L) = -\frac{\partial S(y_S, y_L)}{\partial y_l}, \quad l = S, L$$

The four cases are introduced as follows:

Case 1: $\gamma_{S,i} = 1, \gamma_{L,i} = 1$

$$\begin{aligned} & \Pr [A_i \in (y_S, y_S + dy_S), B_i \in (y_L, y_L + dy_L), \gamma_{S,i} = 1, \gamma_{L,i} = 1] \\ &= \Pr [y_{S,i} \in (y_S, y_S + dy_S), y_{L,i} \in (y_L, y_L + dy_L), C_i > y_L] \\ &= f(y_S, y_L) G(y_L) dy_S dy_L, \end{aligned}$$

Case 2: $\gamma_{S,i} = 1, \gamma_{L,i} = 0$

$$\begin{aligned} & \Pr [A_i \in (y_S, y_S + dy_S), B_i \in (y_L, y_L + dy_L), \gamma_{S,i} = 1, \gamma_{L,i} = 0] \\ &= \Pr [y_{S,i} \in (y_S, y_S + dy_S), C_i \in (y_L, y_L + dy_L), y_{L,i} > y_L] \\ &= \dot{S}_S(y_S, y_L) g(y_L) dy_S dy_L, \end{aligned}$$

Case 3: $\gamma_{S,i} = 0, \gamma_{L,i} = 1$

$$\begin{aligned} & \Pr [A_i \in (y, y + dy), B_i \in (y, y + dy), \gamma_{S,i} = 0, \gamma_{L,i} = 1] \\ &= \Pr [y_{S,i} > t, y_{L,i} \in (y, y + dy), C_i > y] = \dot{S}_L(y, y) G(y) dy, \end{aligned}$$

Case 4: $\gamma_{S,i} = 0, \gamma_{L,i} = 0$

$$\begin{aligned} & \Pr [A_i \in (y, y + dy), B_i \in (y, y + dy), \gamma_{S,i} = 0, \gamma_{L,i} = 0] \\ &= \Pr [y_{S,i} > t, y_{L,i} > y, C_i \in (y, y + dy)] = S(y, y) g(y) dy \end{aligned}$$

Specifically, time y is employed instead of y_L in cases 3 and 4 because we are constrained to observe only a singular time for these two cases. This constraint is imposed by the semi-competing risks nature of the data. Assume non-informative censoring for $y_{S,i}, y_{L,i}$ (the two TTE outcomes of patient i , $i = 1, \dots, n$). On the basis of the aforementioned semi-competing risks framework, the log-likelihood for patients i can be shown as:

$$\begin{aligned} \ell(y_{S,i}, y_{L,i}; \kappa_j) &= \gamma_S \gamma_L \log [f(y_{S,i}, y_{L,i}; \kappa_j)] + \gamma_S (1 - \gamma_L) \log [\dot{S}_S(y_{S,i}, y_{L,i}; \kappa_j)] \\ &\quad + \gamma_L (1 - \gamma_S) \log [\dot{S}_L(y_{S,i}, y_{L,i}; \kappa_j)] + (1 - \gamma_S) (1 - \gamma_L) \log [S(y_{S,i}, y_{L,i}; \kappa_j)] \end{aligned} \tag{6}$$

The cluster-specific parameters κ_j can be updated using the maximum likelihood method for computation convenience based on all the patients in cluster j in iterations. The joint pdf of two TTE outcomes in cluster j can be derived with the cluster-specific parameter κ_j :

$$f(y_{S,i}, y_{L,i} | \lambda, \kappa_j) = \frac{\psi[\Phi^{-1}\{F_S(y_{S,i} | \lambda_{Sj})\}, \Phi^{-1}\{F_L(y_{L,i} | \lambda_{Lj})\} | \kappa_j]}{\phi[\Phi^{-1}\{F_S(y_{S,i} | \lambda_{Sj})\}] \phi[\Phi^{-1}\{F_L(y_{L,i} | \lambda_{Lj})\}]} f_S(y_{S,i} | \lambda_{Sj}) f_L(y_{L,i} | \lambda_{Lj}) \tag{7}$$

where $\phi(\cdot)$ is the standard normal pdf, $\psi(\cdot)$ is the standard bivariate normal pdf with correlation as κ_j , and F_S and F_L are the marginal CDFs for $y_{S,i}$ and $y_{L,i}$, respectively; $f(y_{S,i})$ and $f(y_{L,i})$ are the marginal density functions.

2.4 Prior set-up and posterior computation

For the cluster-specific marginal survival parameters $\lambda_j = \{\lambda_{Sj}, \lambda_{Lj}\}$, we assume the conjugate gamma priors as $Ga(\alpha, \beta)$, where $\alpha = \{\alpha_S, \alpha_L\}$ and $\beta = \{\beta_S, \beta_L\}$ for the two TTE outcomes, respectively:

$$\begin{aligned}\lambda_j &\sim Ga(\alpha, \beta) \\ \alpha, \beta &\propto \pi(\cdot)\end{aligned}$$

where the hyper-prior for parameter α, β is denoted as $\pi(\cdot)$. We define flat gamma priors for parameters α and β , and the prior distributions can be shown as

$$\begin{aligned}\pi(\alpha) &\propto Ga(a_S, b_S) \\ \pi(\beta) &\propto Ga(a_L, b_L)\end{aligned}$$

Therefore, given the specific values for a_S, b_S and a_L, b_L , the target distribution can be obtained by

$$\pi(\lambda_j, \alpha, \beta | \mathbf{Y}) \propto L(\lambda_j | \mathbf{Y}) \pi(\lambda | \alpha, \beta) \pi(\alpha) \pi(\beta)$$

On the basis of the equations (5) and (6), the full conditional distribution can be derived as $\pi(\lambda_j | \mathbf{Y}, \alpha, \beta)$, $\pi(\alpha | \mathbf{Y}, \lambda_j, \beta)$, and $\pi(\beta | \mathbf{Y}, \lambda_j, \alpha)$ to estimate parameter λ_j by performing Gibbs sampling with the Metropolis-Hastings algorithm. Details are provided in the supplementary materials.

2.5 Predictive inference on partition

The adoption of the default choice of $g(\mathbf{x}_j^*)$, as described by [6], could simplify the posterior simulation. The posterior inference in product partition models (2) and (4) is computational identical to the posterior inference that we would obtain if x_i were part of y_i . The auxiliary model $q(\cdot)$ can be extended as

$$\begin{aligned}q(\mathbf{Y}, \mathbf{X} | \rho_n, \lambda_j) &= \prod_{j=1}^{k_n} \prod_{i \in S_j} p(y_{S,i}, y_{L,i} | x_i, \kappa_j, \lambda_j, \eta) q(x_i | \xi_j) \\ q(\lambda, \xi | \eta) &= \prod_{j=1}^{k_n} p(\lambda_j | \eta) q(\xi_j)\end{aligned}$$

The posterior distribution $q(\rho_n | \mathbf{Y}, \mathbf{X})$ under the auxiliary model is identical to $p(\rho_n | \mathbf{Y}, \mathbf{X})$ under the proposed DbTTE-PPMx (5). Specifically,

$p(y_{S,i}, y_{L,i} | x_i, \kappa_j, \lambda_j)$ is illustrated in a copula form, as introduced in section 2.3. To report $p(y_{S,n+1}, y_{L,n+1} | \mathbf{Y}, \mathbf{X}, x_{n+1})$, let $\tilde{x} = \{x_{n+1}\}$ and $\tilde{\mathbf{y}} = \{y_{S,n+1}, y_{L,n+1}\}$ to simplify the notation and let $\tilde{e} = \{e_{n+1}\}$ denote the membership indicator of patient $n+1$; we then find

$$p(\tilde{\mathbf{y}} | \tilde{x}, \mathbf{Y}, \mathbf{X}) = \int p(\tilde{\mathbf{y}} | \tilde{x}, \rho_{n+1}, \mathbf{Y}, \mathbf{X}) dp(\rho_{n+1} | \tilde{x}, \mathbf{Y}, \mathbf{X})$$

The integration equals to summation across all configurations of ρ_{n+1} . By using an importance sampling re-weighting step, we could transfer the form as a posterior integral with respect to $p(\rho_{n+1} | \tilde{x}, \mathbf{Y}, \mathbf{X})$. Recall that $\mathbf{x}_\ell^*$ and $\mathbf{y}_\ell^*$ denote the covariates and two TTE outcomes arranged by cluster ℓ , respectively. Defining $g(\emptyset) = c(\emptyset) \equiv 1$, the prior on $\rho_{n+1} = (\rho_n, \tilde{e})$ can be shown as

$$\begin{aligned} p(\tilde{e} = \ell, \rho_n | \tilde{x}, \mathbf{X}) &\propto \prod_{j \neq \ell} c(S_j) g(\mathbf{x}_j^*) c(S_\ell \cup \{n+1\}) g(\mathbf{x}_\ell^*, \tilde{x}) \\ &= p(\rho_n | \mathbf{X}) \frac{g(\mathbf{x}_\ell^*, \tilde{x}) c(S_\ell \cup \{n+1\})}{g(\mathbf{x}_\ell^*) c(S_\ell)}. \end{aligned} \quad (8)$$

Defining $q_\ell(\tilde{x} | \mathbf{x}_\ell^*) \equiv g(\mathbf{x}_\ell^*, \tilde{x}) / g(\mathbf{x}_\ell^*)$, the posterior predictive distribution can be written as

$$\begin{aligned} p(\tilde{\mathbf{y}} | \tilde{x}, \mathbf{Y}, \mathbf{X}) &\propto \int_{\ell=1}^{k_n+1} p(\tilde{\mathbf{y}} | \tilde{x}, \mathbf{y}_\ell^*, \mathbf{x}_\ell^*, \tilde{e} = \ell) q_\ell(\tilde{x} | \mathbf{x}_\ell^*) \\ &\quad \times \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)} p(\rho_n | \mathbf{Y}, \mathbf{X}) d\rho_n \end{aligned} \quad (9)$$

For each of these estimated ρ_n , the $\tilde{e} = l$ is generated with probabilities proportional to $w_\ell = q_\ell(\tilde{x} | \mathbf{x}_\ell^*) c(S_l \cup \{n+1\}) / c(S_l)$ and $\tilde{\mathbf{y}}$ is generated from $p(\tilde{\mathbf{y}} | \mathbf{y}_\ell^*, \tilde{e} = \ell)$, weighted with w_ℓ . In a specific scenario where $l = k_n + 1$, we define the weight w_{k_n+1} as $w_{k_n+1} = g(\tilde{x}) c(n+1)$. To report the posterior inference for the random partition, we will introduce a two-stage approach in the next subsection.

2.6 Posterior inference in two stages

At stage 1, the posterior simulation scheme is performed. In each iteration of the posterior simulation scheme, the partition ρ_n and number of cluster K are obtained. The optimal number $K = k$ of clusters can be determined on the basis of the maximum posteriori after a period of iterations. At stage 2, we aim to summarize the posterior samples to derive the point estimate of the clustering. We initialize stage 2 by retaining all the parameters and cluster structures obtained from the final iteration of stage 1. These parameters serve as the initial values for stage 2. Building upon the optimal number of clusters we obtained from stage 1, we subsequently fix the number of clusters to perform the MCMC. That is, the MCMC process is continued while conditioned on $K = k$ and on the acquisition of posterior MCMC samples of the random partition. Let P define an $(n \times n)$ adjacency matrix, with $P[ij] = 1$ if patients i and j are in the same cluster. We obtain P^m , where $m = 1, \dots, M$ of adjacency matrices, i.e., ***

$P^m \sim p(P|k^*, data)$. Finally, we use the method proposed by Dahl [19] (hereafter referred to as the least-squares method) to summarize the MCMC posterior samples.

In the context of applying the least squares method, we use the Hemming distance as a distance measure, denoted by $\delta(P_S, P_L)$. Subsequently, a point summary of random partition is determined as $P^* = \operatorname{argmin}_Q E(\delta(P, Q)|data)$, with the $E(\cdot)$ being the posterior expectation over the random P . The determination of this optimal summary is achieved through the minimization of summary Q . To compute the posterior expectation, we use the formula:

$$E(\delta(P, Q)|data) = 1/M \left(\sum_{m=1}^M \delta(P^m, Q) \right)$$

Finally, the optimal value is evaluated as a minimum over $Q \in P^m$, where $m = 1, \dots, M$.

The R-package for DbTTE-PPMx is available at the website on <https://biomedstat.med.kyoto-u.ac.jp/statpkg.html>. In addition, an example of the implementation of DbTTE-PPMx is presented in the Supplementary Materials.

3 Simulation study

We conducted a simulation study to evaluate the performance of PPMx and DbTTE-PPMx in the presence of multiple combinations of covariates. The simulation study was designed as a nonlinear regression method applied to different covariate combinations, which is the same method used by [6], enabling a comparison. We assume three scenarios for covariates. In the first scenario (Scenario 1C2B), there are one continuous covariate and two binary covariates – thus, three covariates in total. In the second scenario (Scenario 1C2B2N), we add two more nuisance binary covariates to 1C2B; 1C2B2N thus contains five covariates in total. Finally, in the third scenario (Scenario 1C2B7N), we add seven nuisance covariates to 1C2B, resulting in ten covariates in total. We formulate the covariate combinations by exclusively considering non- nuisance covariates. We trichotomize the continuous covariates into three levels as $x_{1,d} \in \{-1, 0, 1\}$ and the binary covariates $x_{2,d}, x_{3,d} \in \{0, 1\}$, where d denotes the indicator of the covariate combinations. The variable $x_{1,d}$ with -1, 0, and 1 represents small, medium, and large levels of continuous covariates, respectively. In total, there are 12 covariate combinations in this simulation study (i.e., $d = 1, \dots, 12$).

3.1 Simulation settings

We generated 100 datasets, each comprising $N = 300$ samples, as the primary pattern. To extend our analysis, we created additional datasets with the sample size $N = \{100, 500\}$. Consequently, datasets with sample size $N = \{100, 300, 500\}$ were generated using the following sampling model. On the basis of the covariates in our application study, we assume a continuous covariate x_1 following a log-normal distribution with mean = $\log(38)$ and standard deviation = 0.6; in addition, two binary covariates x_2 and x_3 follow a Bernoulli distribution with probabilities p_2 and p_3 , respectively, where $p_2 = 0.3$ for $x_2 = 1$ and 0 otherwise and $p_3 = 0.6$ for $x_3 = 1$ and

0 otherwise. For TTE outcomes, two correlated random variables $s_{S,i}$ and $s_{L,i}$ with uniform distributions on $(0,1)$ will be initially generated from a bivariate Gaussian copula with a specified correlation matrix for each patient i . The correlation matrix is defined as

$$\Sigma = \begin{bmatrix} 1 & h \\ h & 1 \end{bmatrix}$$

Recall that $y_{S,i}$ and $y_{L,i}$ denote the two TTE outcomes for patient i with the constraint that $y_{S,i} \leq y_{L,i}$. Let $Y_S = \{y_{S,1}, \dots, y_{S,n}\}$ and $Y_L = \{y_{L,1}, \dots, y_{L,n}\}$ denote the entire set of two TTE outcomes with constant hazard following a marginal exponential distribution [18]; $y_{S,i}$ and $y_{L,i}$ can be obtained by using the marginal inverse survival function of S_S^{-1} and S_L^{-1} , respectively:

$$y_{S,i} = S_S^{-1}(s_{S,i}) \text{ and } y_{L,i} = S_L^{-1}(s_{L,i})$$

To obtain the marginal inverse survival function, the median $y_{S,i}$ and $y_{L,i}$, where patient i belongs to covariate combinations d , are assumed as $\theta = \{\theta_{Sd}, \theta_{Ld}\}$ which have a linear regression with covariates according to Zippin and Armitage [20].

$$\begin{aligned} y_{S,i} &\sim \text{Exp}\{\log(2)/\theta_{Sd}\} \\ y_{L,i} &\sim \text{Exp}\{\log(2)/\theta_{Ld}\} \\ \theta_{Sd}^{-1} &= \beta_1 \mathbf{x}_{m,d} + c_1 \\ \theta_{Ld}^{-1} &= \beta_2 \mathbf{x}_{m,d} + c_2 \end{aligned}$$

$$\begin{aligned} \beta_1 \mathbf{x}_{m,d} &= \begin{bmatrix} x_{1,1} & x_{2,1} & x_{3,1} \\ x_{1,2} & x_{2,2} & x_{3,2} \\ \dots & \dots & \dots \\ x_{1,k} & x_{2,k} & x_{3,k} \end{bmatrix} \cdot \begin{bmatrix} \beta_{S,1} \\ \beta_{S,2} \\ \beta_{S,3} \end{bmatrix} \\ \beta_2 \mathbf{x}_{m,d} &= \begin{bmatrix} x_{1,1} & x_{2,1} & x_{3,1} \\ x_{1,2} & x_{2,2} & x_{3,2} \\ \dots & \dots & \dots \\ x_{1,k} & x_{2,k} & x_{3,k} \end{bmatrix} \cdot \begin{bmatrix} \beta_{L,1} \\ \beta_{L,2} \\ \beta_{L,3} \end{bmatrix} \end{aligned}$$

where $\beta_1 = \{\beta_{S,1}, \beta_{S,2}, \beta_{S,3}\}$ and $\beta_2 = \{\beta_{L,1}, \beta_{L,2}, \beta_{L,3}\}$ are vectors of regression parameters for two TTE outcomes, respectively; $\mathbf{x}_{m,d}$ is a matrix for covariate combinations, where $m = \{1, 2, 3\}$ and $d = 1, \dots, 12$. For θ_{Sd} , $\beta_1 = \{0.2, -0.6, -0.5\}$ and constant $c_1 = 15$; for θ_{Ld} , $\beta_2 = \{0.8, -0.4, -0.1\}$ and constant $c_2 = 21$. There is an interaction of $x_{2,d}$ and $x_{3,d}$ with a shorter survival for $(x_{2,d}, x_{3,d}) = (1, 1)$. With the above settings, the regression parameters vary across different clusters due to the different combinations of covariates. Independent, non-informative censoring time is assumed to follow a uniform distribution; that is, $C \sim U[0, 200]$. We obtain $y_{S,i}^* = y_{S,i} \wedge y_{L,i} \wedge C$, $\gamma_{S,i} = I(y_{S,i} \leq y_{L,i} \wedge C)$ and $y_{L,i}^* = y_{L,i} \wedge C$, $\gamma_{L,i} = I(y_{L,i} \leq C)$. For each

correlation coefficient $\kappa \in \{0, 0.25, 0.5, 0.75\}$, 100 datasets will be sampled using Gaussian copula, resulting in different censoring rates: $\{5\%, 29\%\}, \{6\%, 27\%\}, \{7\%, 25\%\}$ and $\{8\%, 23\%\}$, respectively. For the other two scenarios, in addition to the three aforementioned covariates, we introduce two additional binary covariates for Scenario 1C2B2N. In addition, to mimic the covariates in the application, we simulate one additional continuous covariate, five binary covariates, and one count covariate for Scenario 1C2B7N. The additional continuous covariate follows a log-normal distribution with a mean equal to 0 and a variance equal to 1, whereas the additional binary covariates are a Bernoulli-distributed with a probability of 0.2. The count covariate follows a Poisson distribution with a mean equal to 1. Importantly, these additional covariates do not affect the TTE outcomes. The number of covariates for these three scenarios is 3, 5, and 10, respectively. The maximum follow-up time is 72 months (based on the largest OS in the application).

To evaluate the predictive performance, the root-mean-square error (RMSE) (10) was used as the criterion. It was computed as the squared error between the predicted median $\hat{Y}_{d,i}, d = S, L$ from both methods and the generated simulation truth $Y_{d,i}$ across the different correlation coefficients and sample sizes.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\hat{Y}_{d,i} - Y_{d,i} \right)^2} \quad (10)$$

3.2 Simulation results

Table 1 provides a summary of the average RMSE across all covariate combinations for different correlation coefficients and sample sizes. As expected, DbTTE-PPMx provided better prediction accuracy across all of the cases, leveraging the advantages of considering two TTE outcomes simultaneously. Overall, we observed a decreasing tendency in the RMSE for both methods, predicting both median Y_S and Y_L as the correlation coefficients increased. Specifically, the RMSE obtained from Y_L is larger than that obtained from Y_S , possibly because of its greater variation.

Table 1: Simulation results for the comparison between the DbTTE-PPMx and PPMx methods, as expressed by the root mean squared error of predicted median Y_S and Y_L under the three covariate scenarios with three total sample size values $\{100, 300, 500\}$ for four correlation coefficients $\{0, 0.25, 0.50, 0.75\}$ between Y_S and Y_L .

Sample size	TTE outcome	Scenario	Method	Correlation			
				0	0.25	0.5	0.75
100	Y_S	1C2B	DbTTE-PPMx	5.30	4.81	4.44	3.75
			PPMx	5.77	5.38	5.47	5.33
		1C2B2N	DbTTE-PPMx	5.41	4.89	4.56	3.94
			PPMx	6.07	5.74	6.22	6.30
		1C2B7N	DbTTE-PPMx	5.58	5.04	4.72	4.20
			PPMx	6.34	6.09	6.65	6.37
	Y_L	1C2B	DbTTE-PPMx	13.95	12.76	11.10	9.13
			PPMx	19.24	18.09	15.86	13.92
		1C2B2N	DbTTE-PPMx	14.09	13.05	11.39	9.40
			PPMx	20.12	18.87	16.69	14.22
		1C2B7N	DbTTE-PPMx	14.51	13.45	12.03	9.72
			PPMx	19.98	18.38	16.69	14.70
300	Y_S	1C2B	DbTTE-PPMx	4.57	3.91	3.30	2.61
			PPMx	5.89	5.30	4.79	4.07
		1C2B2N	DbTTE-PPMx	4.88	4.24	3.70	2.94
			PPMx	6.06	5.51	5.17	4.79
		1C2B7N	DbTTE-PPMx	5.50	4.96	4.39	3.77
			PPMx	6.40	6.07	5.60	5.21
	Y_L	1C2B	DbTTE-PPMx	13.20	11.80	10.40	8.07
			PPMx	16.60	15.00	13.10	9.65
		1C2B2N	DbTTE-PPMx	13.40	12.00	10.70	8.25
			PPMx	16.60	14.90	13.10	9.87
		1C2B7N	DbTTE-PPMx	13.80	12.40	11.10	8.90
			PPMx	19.50	17.80	15.30	12.30
500	Y_S	1C2B	DbTTE-PPMx	4.36	3.81	3.21	2.38
			PPMx	5.94	5.48	4.80	3.93
		1C2B2N	DbTTE-PPMx	4.53	4.04	3.40	2.65
			PPMx	6.09	5.61	5.06	4.29
		1C2B7N	DbTTE-PPMx	5.30	4.83	4.12	3.23
			PPMx	6.27	5.99	5.55	5.27
	Y_L	1C2B	DbTTE-PPMx	13.15	11.70	10.03	8.06
			PPMx	15.80	13.74	11.34	8.84
		1C2B2N	DbTTE-PPMx	13.40	11.88	10.09	8.19
			PPMx	15.70	13.80	11.33	9.25
		1C2B7N	DbTTE-PPMx	13.69	12.40	10.73	8.51
			PPMx	18.76	16.45	14.07	11.04

Focusing on the simulations with $N = 300$, we find that, under Scenario 1C2B, DbTTE-PPMx consistently outperformed PPMx in terms of a smaller RMSE for predicting both median Y_S and Y_L and a more significant decrease for predicting median Y_S across all correlation coefficients. We observed that, when the correlation coefficients are 0.75, DbTTE-PPMx is modestly more favorable than PPMx, particularly in the prediction of the median Y_S , underscoring its capability in capturing the intricate relationships between the two TTE outcomes. That is, the advantage of DbTTE-PPMx became more pronounced with increasing correlation coefficients, as reflected by a decrease in the RMSE. Under Scenario 1C2B2N with two nuisance covariates, the increase in RMSE for predicting both median Y_S and Y_L was not substantial for either method across all correlation coefficients. DbTTE-PPMx continued to achieve a smaller RMSE than PPMx across all of the correlation coefficients. Under Scenario 1C2B7N, with seven nuisance covariates, the RMSE for predicting both median Y_S and Y_L was increased for both methods across all of the correlation coefficients. Notably, the increase in the RMSE was more pronounced for PPMx than for DbTTE-PPMx particularly in the prediction of median Y_L . This greater increase in the RMSE for PPMx might indicate the benefit of the joint analysis of two TTE outcomes that enhances the robustness of a large number of nuisance covariates, thereby improving the predictive performance of TTE outcomes.

The simulation results for different sample sizes are also summarized in Table A1. With $N = 100$, whereas PPMx exhibited erratic RMSE patterns, DbTTE-PPMx maintained the tendency of the RMSE to decrease with increasing correlation coefficients. Compared with the results corresponding to $N = 300$, a slight increase in the RMSE was observed; DbTTE-PPMx continued to outperform PPMx, suggesting its capacity to deliver a more reliable prediction with a smaller sample size than PPMx. When the sample size was increased to $N = 500$, both methods continued to show a decreasing RMSE, with DbTTE-PPMx consistently performing better, particularly with higher correlation coefficients. Compared with the results corresponding to $N = 300$, the RMSE obtained from both methods for both TTE outcomes decreased slightly. In particular, the improvement in RMSE for predicting the median Y_L by PPMx was more noticeable than that for predicting the median Y_S .

The RMSEs of the predicted median Y_S and Y_L for 12 different covariate combinations across different scenarios and sample sizes are summarized in Tables S1-S9 in the supplementary materials. Across the three different scenarios, we still observed that DbTTE-PPMx led to a smaller RMSE than PPMx for the majority of covariate combinations in predicting both median Y_S and Y_L , showing a consistent tendency.

In summary, we found that the proposed DbTTE-PPMx model led to a smaller RMSE compared with PPMx, particularly under the condition of higher correlation coefficients. This observation suggests that, in situations involving two TTE outcomes with higher correlation coefficients, the DbTTE-PPMx model can potentially assist in the medical decision-making process by providing more reliable predictions of TTE outcomes compared with PPMx. In addition, an increase in sample size enhanced the prediction of the median TTE outcomes for both methods but more notably for PPMx. However, DbTTE-PPMx consistently demonstrated smaller RMSE values than PPMx,

which might highlight the capability of DbTTE-PPMx to provide reliable predictions even with small sample sizes.

4 Application

In the application study, we conducted data processing and analysis on a breast cancer clinical trial dataset comprising 278 patients [21]. The dataset includes two TTE outcomes measured in months: TTF and OS with eight baseline covariates including two continuous covariates, five binary covariates and one count covariate. In this study, OS is defined as the duration from the commencement of the first set of chemotherapy until the death from any cause. TTF is defined as the duration from the initiation of chemotherapy to discontinuation of chemotherapy for any reason, e.g., disease progression, treatment toxicity. The continuous covariates, there are age in years at baseline and disease-free interval (DFI) defined as the time interval in month from the initiation of primary therapy to the diagnosis of recurrence. The binary covariates include two treatment indicators: one indicating if the patient received paclitaxel chemotherapy and bevacizumab (1 for received, 0 for not received) and another one indicating the receipt of eribulin (1 for received, 0 for not received). The binary covariates also include a metastasis indicator: denoting the presence (1) or absence (0) of liver metastasis at the time of diagnosis; a breast cancer stage indicator, with 1 denoting stage IV and with 0 denoting recurrence; and a breast cancer subtype indicator, with 0 representing the luminal subtype and with 1 representing the triple-negative breast cancer (TNBC) subtype. Finally, for the count covariates, the number of metastases: it represents the total number of metastases. The applied DbTTE-PPMx model is constructed by combining models (3) and (5), as introduced in section 2. To model the eight covariates, let $x_{i,p} = (x_{i,1}, \dots, x_{i,8})$ and the similarity function can be shown as $g(\mathbf{x}_j^*) = \prod_{p=1}^8 g^p(x_{j,p}^*)$. Following the suggestions described in section 2.2, we specified the similarity function for each type of covariate accordingly. Specifically, continuous covariates are standardized to have a sample mean of zero and a unit standard deviation. The fixed parameters of similarity function are defined as $m = 0, \vartheta = 0.25, B = 1$ and $D_0 = 1$.; For categorical covariates, the Dirichlet parameters are defined as $\alpha = (0.1, 0.1)$. For count covariate, the shape and scale parameters of Gamma prior are defined as $(a, b) = (1, .025)$. To clarify the characteristics of our proposed DbTTE-PPMx, we used the two-stage approach introduced in section 2.6 to perform clustering to obtain the cluster structure of patients. The two-stage framework was both applied to PPMx and the proposed DbTTE-PPMx. The results obtained from stage 1 are shown in Figure 1. This illustrates the frequency of the number of clusters obtained from DbTTE-PPMx, which is in between the results obtained from PPMx by analyzing TTF and OS separately. Stage 2 is conducted using nine clusters, as shown in Figure 1.

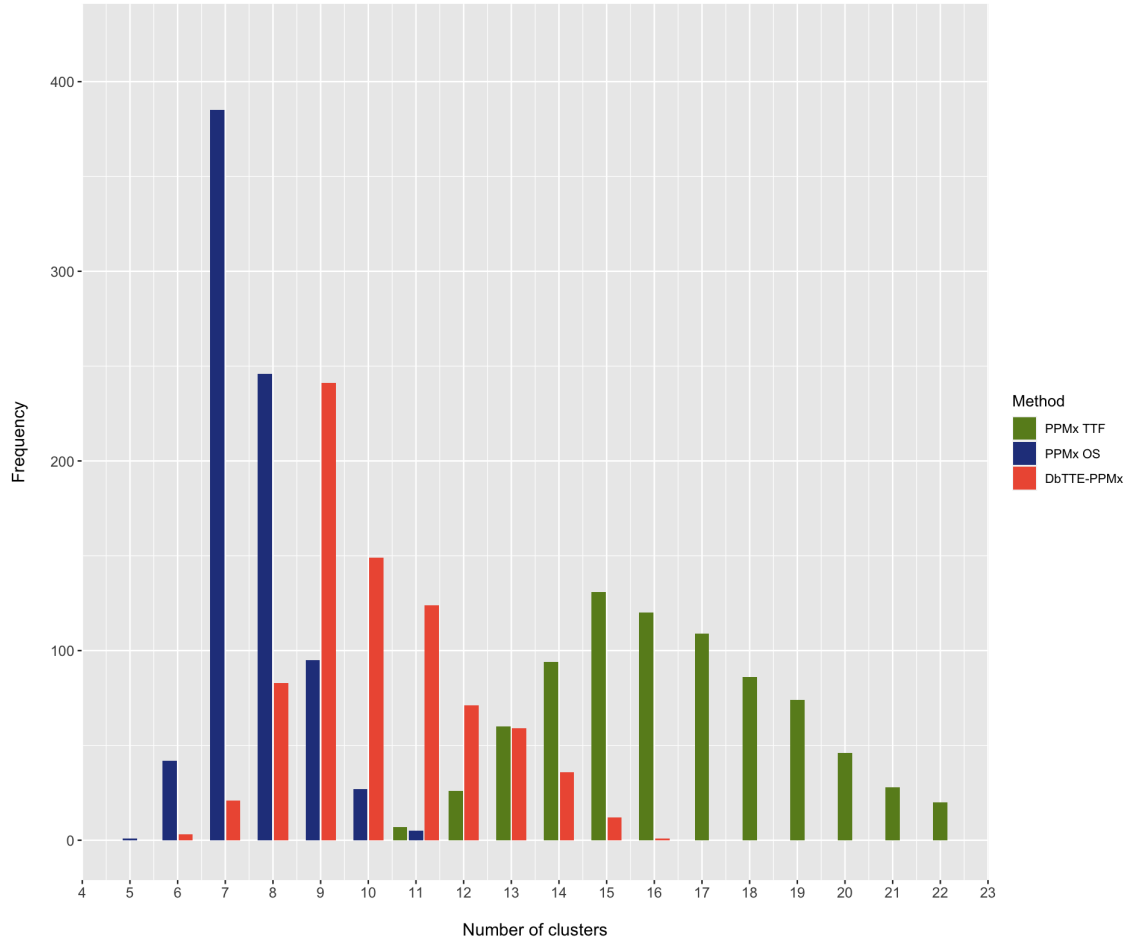
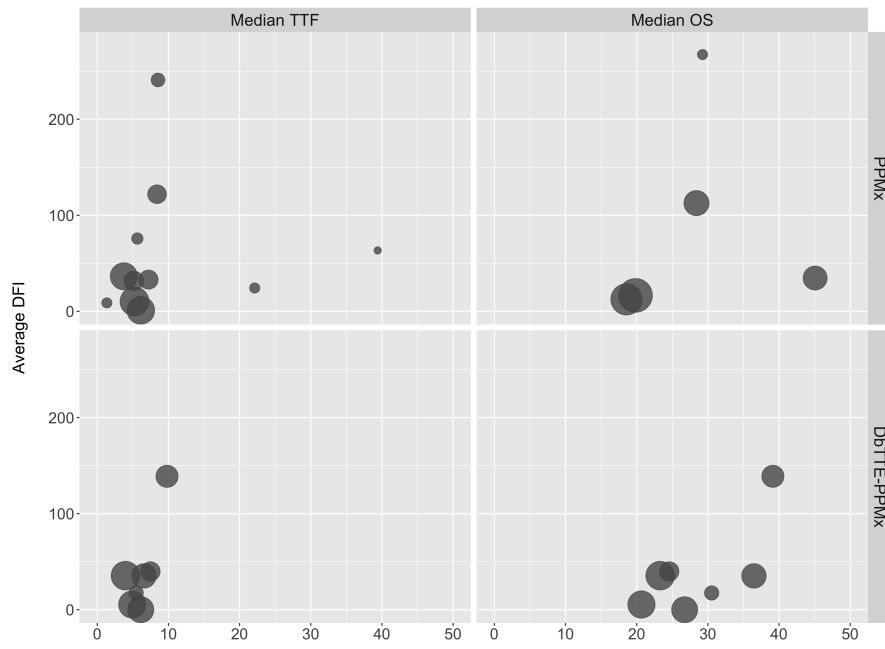
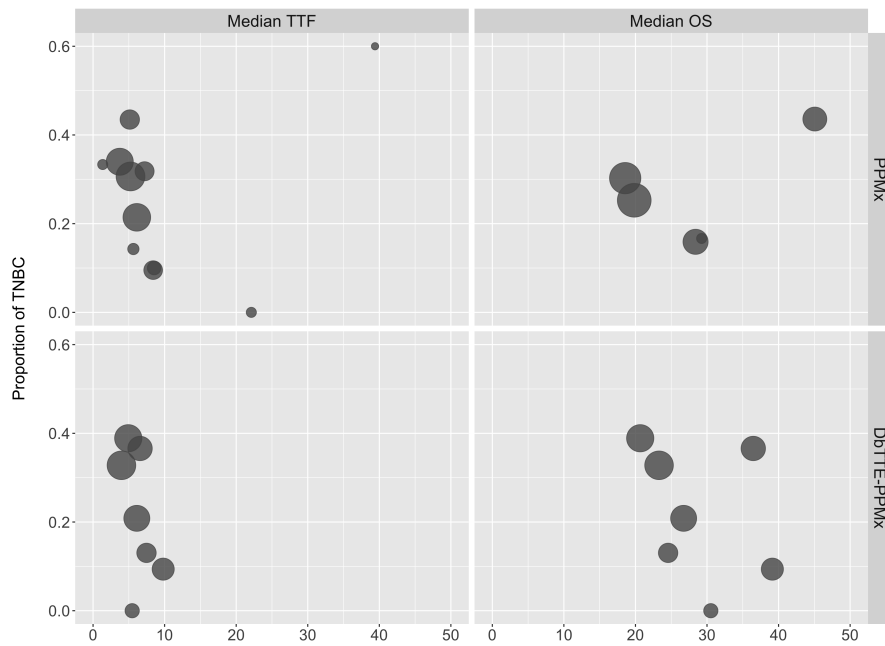


Fig. 1: Frequency of the number of clusters obtained using PPMx and DbTTE-PPMx to analyze TTF and OS



(a) Median TTF and OS with DFI



(b) Median TTF and OS with TNBC

Fig. 2: Summary of the median survival time and the average disease-free interval (DFI) in (a) and the proportion of triple-negative breast cancer (TNBC) in (b) per cluster, as generated using PPMx and DbTTE-PPMx. The size of the bullets is relative to the average cluster size.

One of the notable advantages of our proposed DbTTE-PPMx is its ability to jointly analyze TTF and OS and interpret the results simultaneously. In the context of the application data, we show the cluster structures based on the covariates DFI and TNBC. The relationship between covariates and TTE outcomes is visualized in Figure 2. The results were obtained from PPMx, which analyzes one TTE outcome at a time. In the bottom-right plot of Figure 2 (a), a tendency is observed; as the average DFI increases, the median OS also exhibits an increase. However, such an interpretable result is challenging to derive from the plots generated by PPMx. In addition, patients with TNBC are widely recognized to experience a rapid progression and shorter OS compared with patients with other subtypes of breast cancer (e.g. luminal subtype) [22, 23]. Figure 2 (b) depicts the relationship between median survival time (TTF and OS) and the proportion of TNBC within each cluster. Notably, we observe a tendency in the bottom-right plot in Figure 2 (b); the OS tends to decrease with an increasing proportion of TNBC. However, the tendency remains undetectable in the PPMx-generated plots, which does not align with prior medical knowledge. Using the information of the cluster structure obtained from the proposed model, it is possible to examine the specific characteristics of each clusters to gain deeper insights. For example, one cluster exhibits the longest OS of 39.1 months and the longest TTF of 9.8 months. The characteristics of this cluster are its notably longest average DFI of 138.9 days and the second lowest proportion of patients with TNBC (9%). In contrast, another cluster has the shortest OS of 20.7 months, achieves the third shortest TTF of 4.9 months. This cluster has the second shortest average DFI of 5.4 days and highest proportion of patients with TNBC (39%). These observations indicate that the cluster with the characteristics of short DFI and a high proportion of patients with TNBC might experience short TTF and OS. Moreover, There is a cluster that has the third shortest average DFI of 17.4 months but the lowest portion of patient with TNBC (0%). This observation suggests that in the absence of TNBC, patients might still have a longer OS after treatment with eribulin chemotherapy , even with a short DFI. Importantly, these insights are not available when modelling the two TTE outcomes separately using PPMx, underscoring the value of our proposed model in capturing the intricate relationships between cluster characteristics and TTE outcomes.

5 Discussion

This study proposes the DbTTE-PPMx model to elucidate the subgroup structure by analyzing patients' covariates, that is, clinical characteristics in conjunction with two TTE outcomes. Our model is an extension of the PPMx, which is the random partition model framework, enhanced by the incorporation of the copula method and the semi-competing risks framework to account for interpretation of the intricate relationship between two TTE outcomes. Leveraging this interpreted relationship might provide the feasibility to perform subgroup identification and subgroup structure interpretation with consideration of the two TTE outcomes. Although we have primarily discussed the application of our methodology to TTF and OS in the Application section, any other pair of TTE outcomes, such as PFS and survival post-progression (SPP) [24], is applicable in our model.

The proposed DbTTE-PPMx model can be used to analyze data from a phase II clinical trial, and then design a subsequent phase III clinical trial. Suppose that a phase III trial will evaluate progression-free survival (PFS) time as the primary endpoint and OS as the key secondary endpoint, as discussed in [25]. On the basis of a cluster structure obtained by analyzing both PFS and OS outcomes in combination with relevant patients' baseline covariates from a phase II trial, it could help to select or prioritize an appropriate target patient population for a phase III trial, in which patients could obtain a sufficient benefit in prolonging not only PFS but also OS. From another perspective, it could also help identify the potential high-risk patients to improve the chance that those recruited patients have an acceptable risk-benefit balance. This could minimize the potential risks to vulnerable patients and lowers the chances of trial failures.

In this manuscript, the proposed DbTTE-PPMx only considered right censoring. However, other types of censoring, such as interval censoring, can occur in actual clinical trial data. Moreover, we have only considered the univariate censoring time, which means that scenarios involving two censoring times have not been explored. More complicated censoring scenarios might be considered in future work.

Acknowledgements. We thank Dr Shoichiro Ohtani and Dr Masahiko Ikeda for providing us with invaluable clinical data. Kang Lei’s work was supported by JST SPRING, grant number JPMJSP2110. Satoshi Morita was partially supported by the Project Promoting Clinical Trials for Development of New Drugs (19lk0201061h0004) from the Japan Agency for Medical Research and Development (AMED). We sincerely appreciate the insightful comments and suggestions of the reviewers and the editor, which have significantly improved the quality of this manuscript.

Declarations

Competing interests

The authors have declared no conflict of interest.

References

- [1] Pazdur, R.: Endpoints for assessing drug activity in clinical trials. *The oncologist* **13**(S2), 19–21 (2008)
- [2] Maroun, J.: The significance of progression-free survival as an endpoint in evaluating the therapeutic value of antineoplastic agents. *Current Oncology* **18**(Suppl 2), 3 (2011)
- [3] Hess, L.M., Brnabic, A., Mason, O., Lee, P., Barker, S.: Relationship between progression-free survival and overall survival in randomized clinical trials of targeted and biologic agents in oncology. *Journal of Cancer* **10**(16), 3717 (2019)
- [4] Shukuya, T., Mori, K., Amann, J.M., Bertino, E.M., Otterson, G.A., Shields, P.G., Morita, S., Carbone, D.P.: Relationship between overall survival and response or progression-free survival in advanced non-small cell lung cancer patients treated with anti-pd-1/pd-l1 antibodies. *Journal of thoracic oncology* **11**(11), 1927–1939 (2016)
- [5] Barry, D., Hartigan, J.A.: Product partition models for change point problems. *The Annals of Statistics*, 260–279 (1992)
- [6] Müller, P., Quintana, F., Rosner, G.L.: A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics* **20**(1), 260–278 (2011)
- [7] Page, G.L., Quintana, F.A., Müller, P.: Clustering and prediction with variable dimension covariates. *Journal of Computational and Graphical Statistics* **31**(2), 466–476 (2022)
- [8] Pedone, M., Argiento, R., Stingo, F.C.: Personalized treatment selection via product partition models with covariates. *Biometrics* **80**(1), 003 (2024)

- [9] Kavelaars, X., Mulder, J., Kaptein, M.: Decision-making with multiple correlated binary outcomes in clinical trials. *Statistical Methods in Medical Research* **29**(11), 3265–3277 (2020)
- [10] Sun, T., Liu, Y., Cook, R.J., Chen, W., Ding, Y.: Copula-based score test for bivariate time-to-event data, with application to a genetic study of amd progression. *Lifetime data analysis* **25**, 546–568 (2019)
- [11] Li, D., Hu, X.J., Wang, R.: Evaluating association between two event times with observations subject to informative censoring. *Journal of the American Statistical Association* **118**(542), 1282–1294 (2023)
- [12] Wei, Y., Wojtyś, M., Sorrell, L., Rowe, P.: Bivariate copula regression models for semi-competing risks. *Statistical Methods in Medical Research* **32**(10), 1902–1918 (2023)
- [13] Fine, J.P., Jiang, H., Chappell, R.: On semi-competing risks data. *Biometrika* **88**(4), 907–919 (2001)
- [14] Xu, J., Kalbfleisch, J.D., Tai, B.: Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics* **66**(3), 716–725 (2010)
- [15] Hartigan, J.A.: Partition models. *Communications in statistics-Theory and methods* **19**(8), 2745–2756 (1990)
- [16] Quintana, F.A., Iglesias, P.L.: Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(2), 557–574 (2003)
- [17] Varadhan, R., Xue, Q.-L., Bandeen-Roche, K.: Semicompeting risks in aging research: methods, issues and needs. *Lifetime data analysis* **20**(4), 538–562 (2014)
- [18] Fu, H., Wang, Y., Liu, J., Kulkarni, P.M., Melemed, A.S.: Joint modeling of progression-free survival and overall survival by a bayesian normal induced copula estimation model. *Statistics in medicine* **32**(2), 240–254 (2013)
- [19] Dahl, D.B.: Model-based clustering for expression data via a dirichlet process mixture model. *Bayesian inference for gene expression and proteomics* **4**, 201–218 (2006)
- [20] Zippin, C., Armitage, P.: Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics*, 665–672 (1966)
- [21] Nakamoto, S., Watanabe, J., Ohtani, S., Morita, S., Ikeda, M.: Bevacizumab as first-line treatment for her2-negative advanced breast cancer: paclitaxel plus bevacizumab versus other chemotherapy. *in vivo* **34**(3), 1377–1386 (2020)

- [22] Dent, R., Trudeau, M., Pritchard, K.I., Hanna, W.M., Kahn, H.K., Sawka, C.A., Lickley, L.A., Rawlinson, E., Sun, P., Narod, S.A.: Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical cancer research* **13**(15), 4429–4434 (2007)
- [23] Yin, L., Duan, J.-J., Bian, X.-W., Yu, S.-c.: Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Research* **22**, 1–13 (2020)
- [24] Broglio, K.R., Berry, D.A.: Detecting an overall survival benefit that is derived from progression-free survival. *JNCI: Journal of the National Cancer Institute* **101**(23), 1642–1649 (2009)
- [25] Rugo, H.S., Bardia, A., Marmé, F., Cortés, J., Schmid, P., Loirat, D., Trédan, O., Ciruelos, E., Dalenc, F., Pardo, P.G., *et al.*: Overall survival with sacituzumab govitecan in hormone receptor-positive and human epidermal growth factor receptor 2-negative metastatic breast cancer (tropics-02): a randomised, open-label, multicentre, phase 3 trial. *The Lancet* **402**(10411), 1423–1433 (2023)