

Analysis of survival data with cure fraction and variable selection: A pseudo-observations approach

Statistical Methods in Medical Research

2022, Vol. 31(11) 2037–2053

© The Author(s) 2022



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09622802221108579

journals.sagepub.com/home/smm

Chien-Lin Su^{1,2,3} , Sy Han Chiou⁴ , Feng-Chang Lin⁵ ,
and Robert W Platt^{1,2}

Abstract

In biomedical studies, survival data with a cure fraction (the proportion of subjects cured of disease) are commonly encountered. The mixture cure and bounded cumulative hazard models are two main types of cure fraction models when analyzing survival data with long-term survivors. In this article, in the framework of the Cox proportional hazards mixture cure model and bounded cumulative hazard model, we propose several estimators utilizing pseudo-observations to assess the effects of covariates on the cure rate and the risk of having the event of interest for survival data with a cure fraction. A variable selection procedure is also presented based on the pseudo-observations using penalized generalized estimating equations for proportional hazards mixture cure and bounded cumulative hazard models. Extensive simulation studies are conducted to examine the proposed methods. The proposed technique is demonstrated through applications to a melanoma study and a dental data set with high-dimensional covariates.

Keywords

Bounded cumulative hazard, Cox proportional hazard, high-dimensional covariates, mixture cure model, penalized generalized estimating equation

1 Introduction

In the time-to-event analyzes, it is usually assumed that all subjects will eventually experience the event of interest if the follow-up period is sufficiently long. However, in many research fields, including biomedical, genetic, and social studies, some subjects may never experience the event of interest in their lifetime. These subjects are referred to as the cured or nonsusceptible subjects. For example, in a melanoma progression study,¹ the cured patients never experience a melanoma relapse after the initial treatment. On the other hand, in a dental study,² the cured subjects are the teeth that underwent proper periodontal treatments and can last a lifetime. In general, it is difficult to identify the cured subjects, but their presence is signaled by a leveling of the Kaplan-Meier (KM) survival curve at the end of the follow-up, for example, Figure 1(a). Standard models do not account for the cure fraction and could lead to biased estimates of the survival of susceptible subjects.³ Even if the cure fraction is accounted for, the dental study posts additional challenges on high-dimensionality. More than 50 predictors relevant to decision making in periodontal treatments are potential risk factors affecting the teeth' survival. Motivated to tackle these emerging and challenging scientific questions, we propose a new

¹Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada

²Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montréal, Québec, Canada

³Peri and Post Approval Studies, Strategic and Scientific Affairs, PPD, part of Thermo Fisher Scientific, Montréal, Québec, Canada

⁴Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, USA

⁵Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA

Corresponding author:

Chien-Lin Su, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada.

Email: marksu740824@gmail.com

estimating procedure using a pseudo-observations approach for the statistical inference on the parameters of interest and extend the proposed methods to regularized regression.

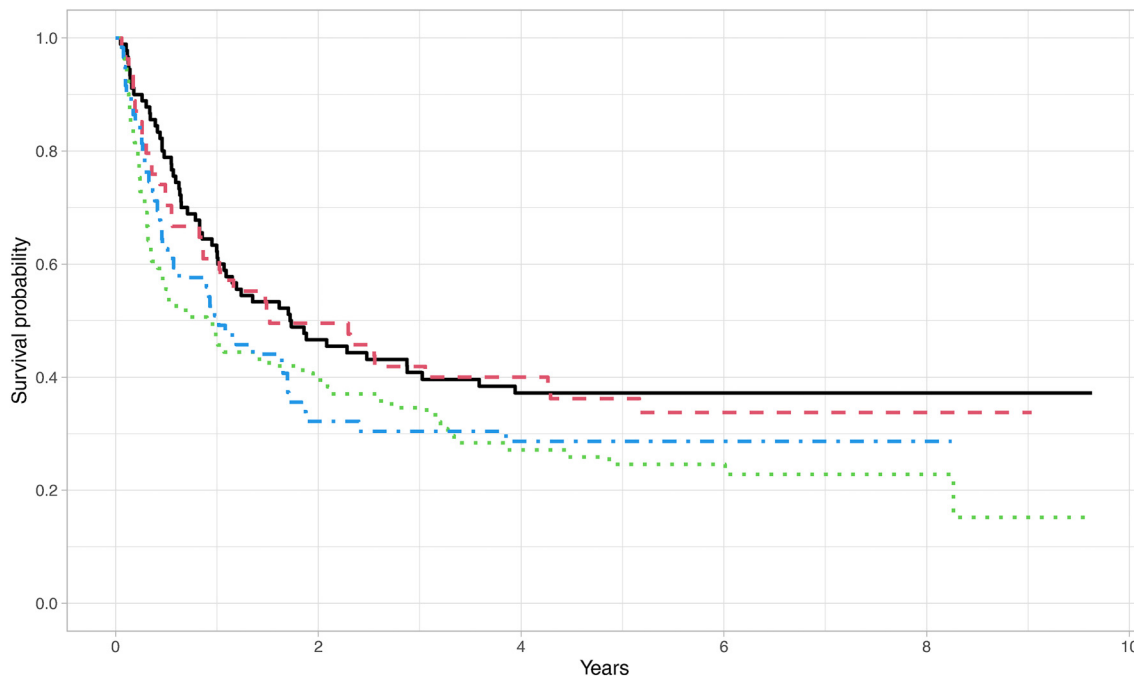
Two types of cure models have been popular in the literature, with most emphasis on the mixture cure (MC) model.⁴ The MC model assumes that the population consists of two types of patients, a cured group in which the patients are not at risk of experiencing the event and a susceptible group in which they eventually experience the event. The MC model consists of two components, *incidence* and *latency*. The former indicates whether the subject is susceptible, and the latter represents the time to the event when the patient is in the susceptible group. For the *incidence* component, a logistic regression model is often used to describe the covariate effects on the cure fraction. In contrast, parametric and semiparametric models have been proposed for the *latency* component to describe the underlying failure time distribution of susceptible subjects. Among those, Weibull model,⁴ generalized F model,⁵ Cox proportional hazards (PH) model,⁶ and accelerated failure time model⁷ have been studied.

The second type of cure model is the bounded cumulative hazard (BCH) model, also known as the promotion time cure model, which models the survival times through an improper survival function, e.g. Tsodikov.⁸ The idea was first introduced by Yakovlev et al.⁹ in biological considerations, in which cancer recurrence is promoted by the number of carcinogenic cells and disease progression. Thus, the parameters specified in the BCH model bear exact biological meaning. Treating the carcinogenic cell counts as latent and nuisance, the BCH model is suitable for any survival data types as long as it is reasonable to assume the data have a cure fraction.¹⁰ Incorporating covariates into the BCH model modifies the cure fraction and introduces a *long-term* covariate effect on the survival. The BCH model has a PH structure through the *long-term* effect parameter. Tsodikov et al.¹¹ further incorporated covariates to the baseline survival function through another PH structure, introducing a *short-term* covariate effect on survival. The two-component BCH model of Tsodikov et al.¹¹ is termed the PHPH model. The MC model and the PHPH model consist of different covariate effects, each providing unique clinical interpretations.

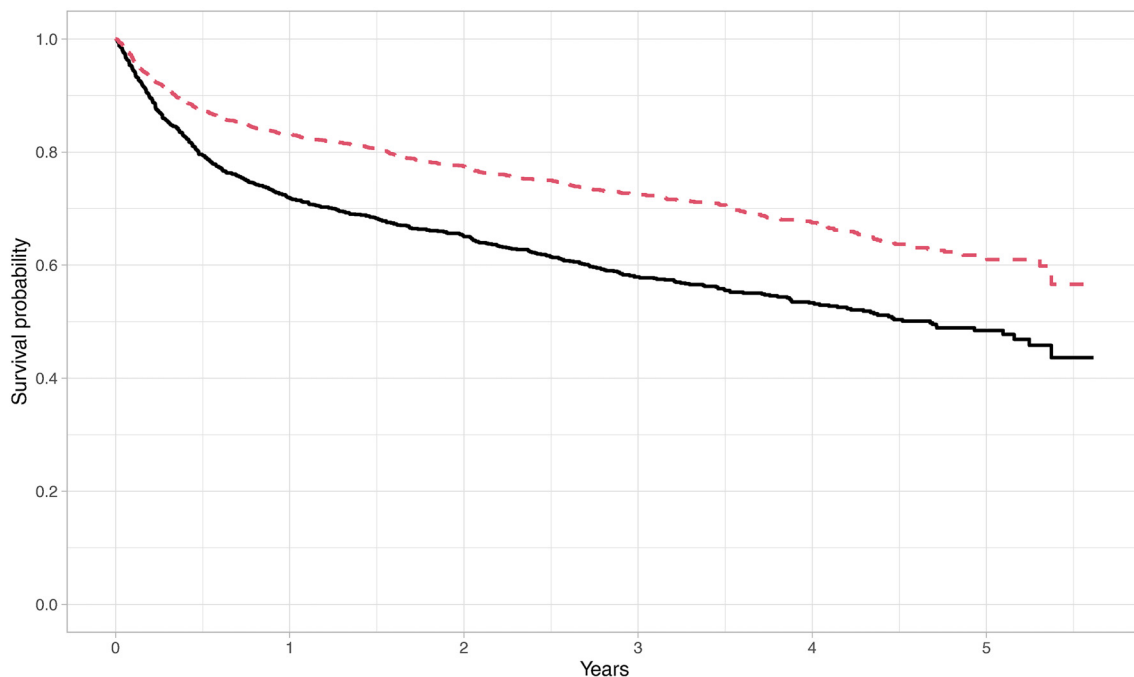
Existing estimating procedures for the MC model and the PHPH model usually involve updating the parameters via expectation-maximization (EM) algorithms to account for latent variables, for example, cure fraction, in the likelihood. However, these approaches could be computationally expensive in high-dimensional data or when the bootstrap method is used in variance estimation. Thus, estimating procedures such as the pseudo-observations approach that does not rely on the EM algorithm are more appealing for practical use. The concept of the pseudo-observations approach is to create pseudo values for the quantities of interest at individual levels using the analogy of leave-one-out cross-validation. These pseudo values are then treated as complete data where standard methods can be conveniently applied. Specifically, for subject $i = 1, \dots, n$, let T_i be independent and identically distributed random variables and X_i be a vector of covariates. The interest lies in modeling $E[f(T_i)|X_i]$, where f is a pre-specified transformation function of T_i . Due to censoring, not all $f(T_i)$ are observed. However, the observed or unobserved $f(T_i)$ can be replaced by its pseudo-observations $\hat{q}^i = n\hat{q} - (n-1)\hat{q}^{-i}$ where \hat{q} is a consistent and (approximately) unbiased estimator for the expectation $q = E[f(T)]$ and \hat{q}^{-i} is the estimator for q using the remaining $n-1$ subjects, leaving subject i out from the sample. The pseudo-observations approach was first proposed by Andersen et al.¹² to model the transition probabilities in multi-state models. Since then, the pseudo-observations approach has been applied to many settings in survival analysis, including survival estimates,¹³ the restricted mean survival times,¹⁴ the cumulative incidence function,¹⁵ the relative survival function,¹⁶ the illness-death model with interval-censored data,¹⁷ and the causal inference for recurrent event data.¹⁸ Large sample properties have also been thoroughly investigated.¹⁹⁻²¹ However, the pseudo-observations approach has not been applied to the analysis of survival data with a cure fraction.

Regularization and variable selection are commonly used in high-dimensional data analysis, but it has been less studied for cure models. Regularized procedures minimize an objective function that consists of a penalty function to reflect sparsity. Some of the popular penalty functions are the least absolute shrinkage and selection operator (LASSO),²² adaptive LASSO (ALASSO),²³ and smoothly clipped absolute deviation (SCAD).²⁴ In the Cox MC models, Liu et al.²⁵ used LASSO and SCAD penalties to select variables based on penalized likelihood functions. Masud et al.²⁶ performed the variable selection based on the ALASSO penalty while considering the linear and nonlinear effects in both components. Masud et al.²⁷ further utilized the ALASSO penalty to the Cox MC and the BCH models. In both works, an EM algorithm was adopted to estimate the parameters, and the bootstrap resampling procedure was used to obtain standard error estimates. Their works can be computationally intensive in high-dimensional data. The penalized BCH model of Masud et al.²⁷ has limited application as it requires additional information on the latent carcinogenic cell counts, making their approach not applicable for survival data without similar biological interpretations. Moreover, the *short-term* covariate effect is not considered in their approach, limiting the understanding of the covariate impact on the timing of disease occurrence.

In this article, we develop new estimating procedures based on the pseudo-observations approach for both the MC and the BCH models. We further extend the proposed method by adopting the penalized generalized estimating equations (PGEE) approach²⁸ to perform variable selection. The proposed work closes the gap on variable selection in cure rate



(a) The Kaplan-Meier survival curve for the melanoma data stratified by treatment and gender: — Treatment/Male; - - Treatment/Female; - - Placebo/Male; ··· Placebo/Female.



(b) The Kaplan-Meier survival curve for the tooth loss data stratified by tooth type: — Non-molar; - - Molar.

Figure 1. The Kaplan-Meier survival curves for the melanoma data and the tooth loss data to access potential cure fraction.

models with pseudo-observations techniques. The proposed approaches are attractive in several aspects. First, pseudo-observations can be straightforwardly used as complete outcomes for the generalized linear model (GLM) without indication of censoring. Second, the proposed estimating procedures are computationally efficient and faster in running time than standard approaches that adopt the EM algorithm for estimation and bootstrapping for standard errors as the unknown regression parameters are estimated via the generalized estimating equations (GEE) approach

with corresponding variance estimates obtained by sandwich estimators. Third, unlike the estimation and variable selection of regression parameters in the *incidence* and *latency* component of the MC model^{6,27} or the *short-term* and *long-term* effects of the BCH model^{11,27} are performed simultaneously within one model, our proposed methods using pseudo-observations can perform the estimation and variable selection separately in each component of the MC model or each effect of the BCH model. Specifically, once the pseudo-observations for the quantity of interest for each component of the MC model or each effect of the BCH model are generated for each subject, they can be modeled with standard methods like GLMs. The GEE and PGEE estimating methods can be applied for parameter estimation and variable selection. Finally, the proposed regression estimators can be easily implemented via standard statistical software.

The remainder of the article is organized as follows. The MC model and the BCH model are reviewed in Section 2. The construction of pseudo-observations is described in Section 3. Inference procedure, model diagnosis, and variable selection are presented in Section 4. Comprehensive simulation results are reported in Section 5. The analysis of two real datasets is provided in Section 6. Concluding remarks are given in Section 7. Asymptotic properties and additional simulation results are provided in the online Supplemental Materials.

2 The cure models

Let Y denote the cure status of a subject, where $Y = 1$ if the subject eventually experiences an event (uncured, susceptible), and $Y = 0$ if the subject is a survivor (cured, non-susceptible). Let $T = YT^* + (1 - Y) \times \infty$ be the survival time, where $T^* < \infty$ is the failure time if the subject is susceptible. In the presence of right censoring, we assume the observed data consist of n independent replicates $(\tilde{T}_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \dots, n$, which are copies of $(\tilde{T}, \delta, \mathbf{X}, \mathbf{Z})$, where $\tilde{T} = \min\{T, C\}$, $\delta = I(T \leq C)$, C is the censoring time, and \mathbf{X} and \mathbf{Z} are vectors of covariates with dimensions p and q , respectively. We allow \mathbf{X} and \mathbf{Z} to be completely distinct, overlapped, or identical. When $\delta = 1$, the subject experienced an event and $Y = 1$. However, when $\delta = 0$, the cure status Y is not observed.

2.1 MC model

The MC model expresses the population survival function as

$$S(t) = (1 - \pi) + \pi S_u(t), \quad (1)$$

where $\pi = P(Y = 1)$ is the uncured rate and $S_u(t)$ is the conditional survival function of T^* given $Y = 1$. The *incidence* component π is assumed to follow a logistic regression model

$$\pi(\mathbf{X}) = P(Y = 1|\mathbf{X}) = \frac{\exp(\alpha_0 + \boldsymbol{\alpha}^\top \mathbf{X})}{1 + \exp(\alpha_0 + \boldsymbol{\alpha}^\top \mathbf{X})}, \quad (2)$$

where α_0 is a scalar and $\boldsymbol{\alpha}$ is a p -column vector. For the *latency* component, we model the conditional survival function via the Cox proportional hazards model

$$\lambda(t|\mathbf{Z}) = \lambda_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{Z}), \quad (3)$$

where $\boldsymbol{\beta}$ is a q -column vector of regression coefficients, and $\lambda_0(t)$ is the unspecified baseline hazard function with the cumulative function $\Lambda_0(t) = \int_0^t \lambda_0(u) du$. Under models (2) and (3), the MC model is called the PHMC model.⁶

2.2 BCH model

Suppose $\Lambda(t)$ is the cumulative hazard function of T^* such that $\Lambda(\infty) = \theta > 0$. Under the BCH model, the population survival function can be written as

$$S(t) = \exp\{-\theta F(t)\}, \quad (4)$$

where $F(t) = \Lambda(t)/\theta$ is a proper cumulative distribution function of a nonnegative random variable with $F(0) = 0$ and $F(\infty) = \Lambda(\infty)/\theta = 1$. As $t \rightarrow \infty$, one has $\lim_{t \rightarrow \infty} S(t) = \exp(-\theta)$ which indicates the cure rate. The covariate effects can be assumed on the impact of the parameter θ with $\theta(\mathbf{X}) \equiv \theta(\mathbf{X}, \gamma_0, \boldsymbol{\gamma})$, where $\theta(\mathbf{X}, \gamma_0, \boldsymbol{\gamma})$ is a known function that relates a $p \times 1$ vector of regression coefficients $\boldsymbol{\gamma}$ to \mathbf{X} with an intercept term γ_0 . This modeling strategy leads to the improper PH model,²⁹ and the covariates have a *long-term* effect because θ describes the long-term survival probability.⁸

A common choice of $\theta(\cdot)$ is the exponential parameterization $\theta(\mathbf{X}) = \exp(\gamma_0 + \boldsymbol{\gamma}^\top \mathbf{X})$. Tsodikov⁸ extends the improper PH model by adding a *short-term* effect by incorporating covariates into $F(t)$ or survival function $\bar{F}(t) = 1 - F(t)$. Specifically, the PHPH model of Tsodikov et al.¹¹ has the form

$$S(t) = \exp[-\theta(\mathbf{X})\{1 - \bar{F}(t)^{\eta(\mathbf{Z})}\}], \quad (5)$$

where $\eta(\mathbf{Z}) = \exp(\boldsymbol{\phi}^\top \mathbf{Z})$ and $\boldsymbol{\phi}$ is a q -column vector of regression coefficients. To avoid overparameterization, we assume the coefficients $\boldsymbol{\phi}$ do not contain an intercept term as suggested in Tsodikov et al.¹¹ When the cure fraction is the only parameter of interest, it makes no difference which model formulation (1) or (4) is chosen to estimate the cure rate nonparametrically. However, the two models become different when additional model assumptions are imposed on the cure rate and the latency distribution of T^* in the MC model.

3 Pseudo-observations

3.1 Pseudo-observations for MC model

A common approach for constructing pseudo-observations is to generate those from a nonparametric estimator of the parameter of interest. The MC model has two parameters, uncured rate π and the conditional survival function $S_u(t)$, to be estimated. There are two candidate estimators for the uncured rate π . The first one is proposed by Maller and Zhou,³⁰ in which the cure rate $1 - \pi$ was estimated by $\hat{S}_{\text{KM}}(t_{\max})$, where $\hat{S}_{\text{KM}}(t)$ is the KM estimator³¹ and t_{\max} is the maximum of the observed event times. The result implies that π can be estimated by $\hat{\pi}_{\text{KM}} = 1 - \hat{S}_{\text{KM}}(t_{\max})$. Following the construction of pseudo-observations in Andersen et al.,¹² one can define the pseudo-observations for subject i by

$$\hat{\pi}_{\text{KM}}^i = n \cdot \hat{\pi}_{\text{KM}} - (n - 1) \cdot \hat{\pi}_{\text{KM}}^{-i}, \quad (6)$$

where $\hat{\pi}_{\text{KM}}^{-i} = 1 - \hat{S}_{\text{KM}}^{-i}(t_{\max})$ is the estimator for π using the remaining $n - 1$ subjects, leaving subject i out from the sample. The second estimator is based on the estimation of θ in Tsodikov³² through the connection between models (1) and (4). To be specific, let $t_{(1)} < t_{(2)} < \dots < t_{(D)}$ be unique observed failure times, and let $t_{(0)} = 0$ and $t_{(D+1)} = \infty$. Let $M_j = \sum_{i=1}^n I(\tilde{T}_i = t_{(j)}, \delta_i = 1)$ and $N_j = \sum_{i=1}^n I(t_{(j)} \leq \tilde{T}_i < t_{(j+1)}, \delta_i = 0)$ be the number of failure times at time $t_{(j)}$ and the number of censored times in the interval $[t_{(j)}, t_{(j+1)})$, respectively. Under model (4), θ can be estimated by $\hat{\theta}_{\text{NP}} = \sum_{k=1}^D \hat{\theta}_k$, where $\hat{\theta}_k = -\log\{(\sum_{\ell=k+1}^D M_\ell + \sum_{\ell=k}^D N_\ell) / (\sum_{\ell=k}^D M_\ell + \sum_{\ell=k}^D N_\ell)\}$, $k = 1, \dots, D - 1$, and $\hat{\theta}_D = -\log(N_D / (N_D + M_D))$. Consequently, $F(t)$ can be estimated by $\hat{F}_{\text{NP}}(t) = \sum_{\{j: t_{(j)} \leq t\}} \hat{J}_j$, where $\hat{J}_j = \hat{\theta}_j / \hat{\theta}_{\text{NP}}$ is the estimated jump size at time $t_{(j)}$. Instead of using the Lagrange multiplier method,³² we use the change of variables approach under the condition $\sum_{k=1}^D J_k = 1$ to obtain $\hat{\theta}_{\text{NP}}$ and $\hat{F}_{\text{NP}}(t)$. The estimating procedure is summarized in Web Appendix A. Since models (1) and (4) have the same cure rate, one could estimate π by $\hat{\pi}_{\text{NP}} = 1 - \exp(-\hat{\theta}_{\text{NP}})$ and create the pseudo-observations for π by

$$\hat{\pi}_{\text{NP}}^i = n \cdot \hat{\pi}_{\text{NP}} - (n - 1) \cdot \hat{\pi}_{\text{NP}}^{-i}, \quad (7)$$

where $\hat{\pi}_{\text{NP}}^{-i}$ is the estimator of π obtained when leaving subject i out from the sample. Web Appendix B shows the behavior of pseudo-observations from (6) and (7) based on simulated data. One can see that these pseudo-observations are not necessary within the range $[0, 1]$.

To create the pseudo-observations for $S_u(t)$, one can express $S_u(t) = \pi^{-1} \cdot (S(t) - (1 - \pi))$ from model (1) and imply that $S_u(t)$ can be estimated by $\hat{S}_{u, \text{KM}}(t) = \{\hat{S}_{\text{KM}}(t) - \hat{S}_{\text{KM}}(t_{\max})\} / \{1 - \hat{S}_{\text{KM}}(t_{\max})\}$. The pseudo-observations for $S_u(t)$ can then be created by

$$\hat{S}_u^i(t) = n \cdot \hat{S}_{u, \text{KM}}(t) - (n - 1) \cdot \hat{S}_{u, \text{KM}}^{-i}(t), \quad (8)$$

where $\hat{S}_{u, \text{KM}}^{-i}(t) = \{\hat{S}_{\text{KM}}^{-i}(t) - \hat{S}_{\text{KM}}^{-i}(t_{\max})\} / \{1 - \hat{S}_{\text{KM}}^{-i}(t_{\max})\}$ is the estimator for $S_u(t)$ when leaving subject i out from the sample.

3.2 Pseudo-observations for BCH model

Under model (4), we propose two approaches to create pseudo-observations for θ and $F(t)$, respectively. Since the cure rate $\lim_{t \rightarrow \infty} S(t) = \exp(-\theta)$ can be nonparametrically estimated by $\hat{S}_{\text{KM}}(t_{\max})$, θ can be estimated by $\hat{\theta}_{\text{KM}} = -\log \hat{S}_{\text{KM}}(t_{\max})$.

Moreover, as mentioned in Section 3.1, θ can also be estimated by $\hat{\theta}_{\text{NP}}$. Thus, the pseudo-observations for θ can be created by one of the following two approaches

$$\hat{\theta}_{\text{KM}}^i = n \cdot \hat{\theta}_{\text{KM}} - (n-1) \cdot \hat{\theta}_{\text{KM}}^{-i}, \quad (9)$$

$$\hat{\theta}_{\text{NP}}^i = n \cdot \hat{\theta}_{\text{NP}} - (n-1) \cdot \hat{\theta}_{\text{NP}}^{-i}. \quad (10)$$

Based on Tsodikov,⁸ $F(t)$ can be consistently estimated by $\hat{F}_{\text{KM}}(t) = \log(\hat{S}_{\text{KM}}(t)) / \log(\hat{S}_{\text{KM}}(t_{\text{max}}))$. As mentioned in Section 3.1, $F(t)$ can also be estimated by $\hat{F}_{\text{NP}}(t) = \sum_{\{j: t_{(j)} \leq t\}} \hat{J}_j$. Thus, the pseudo-observations for $F(t)$ can be created by one of the following two approaches,

$$\hat{F}_{\text{KM}}^i(t) = n \cdot \hat{F}_{\text{KM}}(t) - (n-1) \cdot \hat{F}_{\text{KM}}^{-i}(t), \quad (11)$$

$$\hat{F}_{\text{NP}}^i(t) = n \cdot \hat{F}_{\text{NP}}(t) - (n-1) \cdot \hat{F}_{\text{NP}}^{-i}(t), \quad (12)$$

where $\hat{F}_{\text{KM}}^{-i}(t)$ and $\hat{F}_{\text{NP}}^{-i}(t)$ are estimators of $F(t)$ when leaving subject i out from the sample.

4 Statistical inference

The statistical inference of the pseudo-observations approach is based on the asymptotic unbiased property of the pseudo-observations for the parameter of interest.³³ Following Jacobsen and Martinussen²⁰ and Overgaard et al.,²¹ we present the proofs of the asymptotic unbiased property for proposed pseudo-observations in Web Appendix C of the online Supplemental Materials.

4.1 Estimation of parameters under MC model

Based on the pseudo-observations in Section 3.1, the parameters $(\alpha_0, \boldsymbol{\alpha}^\top)$ in the *incidence* component and $\boldsymbol{\beta}$ in the *latency* component of the PHMC model with (2) and (3) can be estimated separately. To estimate $(\alpha_0, \boldsymbol{\alpha}^\top)$, we consider the following GLM

$$g_1(E[Y_i | \mathbf{X}_i]) = \alpha_0 + \boldsymbol{\alpha}^\top \mathbf{X}_i, \quad (13)$$

where $g_1(x) = \log\{x/(1-x)\}$ is the logit link function for a binary variable. The parameters $(\alpha_0, \boldsymbol{\alpha})$ can be estimated based on the GEE approach³⁴ using pseudo-observations $\hat{\pi}_i, i = 1, \dots, n$ by solving the estimating equations

$$\mathbf{U}(\alpha_0, \boldsymbol{\alpha}) = \sum_{i=1}^n \frac{\partial g_1^{-1}(\alpha_0 + \boldsymbol{\alpha}^\top \mathbf{X}_i)}{\partial (\alpha_0, \boldsymbol{\alpha})} V_{1,i}^{-1} (\hat{\pi}_i - g_1^{-1}(\alpha_0 + \boldsymbol{\alpha}^\top \mathbf{X}_i)) = 0, \quad (14)$$

where $\hat{\pi}^i$ is the pseudo-observations for π , and $V_{1,i}$ is the working variance. Let $(\hat{\alpha}_{0,\text{KM}}^{\text{PO}}, \hat{\boldsymbol{\alpha}}_{\text{KM}}^{\text{PO}})$ and $(\hat{\alpha}_{0,\text{NP}}^{\text{PO}}, \hat{\boldsymbol{\alpha}}_{\text{NP}}^{\text{PO}})$ be the estimators from (14) as $\hat{\pi}^i$ is replaced by $\hat{\pi}_{\text{KM}}^i$ and $\hat{\pi}_{\text{NP}}^i$, respectively.

To estimate $\boldsymbol{\beta}$, the pseudo-observations for $S_u(t)$ are evaluated at several time points and used as responses in the GLM for the covariate effects. Specifically, let $\mathbf{t} = \{t_1, \dots, t_H\}$ be a set of distinct times between 0 and the maximum of observed event time, and let $\hat{S}_u^i(t_h)$ be the pseudo-observations (8) for subject i at time t_h for $h = 1, \dots, H$. We assume the GLM with

$$g_2(E[I(T_i^* > t_h) | \mathbf{Z}_i]) = \xi_{t_h} + \boldsymbol{\beta}^\top \mathbf{Z}_i, \quad (15)$$

where ξ_{t_h} is the intercept at time t_h , $\boldsymbol{\beta}$ is the regression parameters, and $g_2(x)$ is a link function. Common choices for g_2 include the log-log function $\log\{-\log(x)\}$ and log function $\log(x)$. Model (15) becomes the Cox PH model (3) when $g_2(x) = \log\{-\log(x)\}$ and $\xi_{t_h} = \log \Lambda_0(t_h)$. We use the following GEE to estimate the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_H = \{\xi_{t_1}, \dots, \xi_{t_H}\}$

$$\mathbf{U}(\boldsymbol{\psi}) = \sum_{i=1}^n \frac{\partial \mathbf{g}_2^{-1}(\mathbf{t}, \boldsymbol{\psi}; \mathbf{Z}_i)}{\partial \boldsymbol{\psi}} V_{2,i}^{-1} (\hat{S}_u^i(\mathbf{t}) - \mathbf{g}_2^{-1}(\mathbf{t}, \boldsymbol{\psi}; \mathbf{Z}_i)) = 0, \quad (16)$$

where $\boldsymbol{\psi} = (\boldsymbol{\xi}_H, \boldsymbol{\beta})$, $\hat{S}_u^i(t) = (\hat{S}_u^i(t_1), \dots, \hat{S}_u^i(t_H))^\top$, $\mathbf{g}_2^{-1}(t, \boldsymbol{\psi}; \mathbf{Z}_i)$ is a $H \times 1$ vector whose j^{th} component equals $\mathbf{g}_2^{-1}(\xi_{t_j} + \boldsymbol{\beta}^\top \mathbf{Z}_i)$, and $\mathbf{V}_{2,i}$ is a $H \times H$ working covariance matrix that accounts for the correlation inherent from the pseudo-observations.¹²We denote $\hat{\boldsymbol{\psi}}^{\text{PO}} = (\hat{\boldsymbol{\xi}}_H^{\text{PO}}, \hat{\boldsymbol{\beta}}^{\text{PO}})$ as the estimators obtained from solving (16).

4.2 Estimation of parameters under BCH model

Under PHPH model (5), the *short-term* and *long-term* covariate effects can be estimated separately. To estimate the *long-term* effect $(\gamma_0, \boldsymbol{\gamma})$, we consider the following GLM

$$g_3(\hat{\theta}_i) = \gamma_0 + \boldsymbol{\gamma}^\top \mathbf{X}_i + \varepsilon_i, \tag{17}$$

where $\varepsilon_i, i = 1, \dots, n$, are independent and identically distributed (i.i.d.) with mean zero, and g_3 is a link function. Possible choices of $g_3(\cdot)$ are the log link function $\log(x)$ and the log-log function $\log\{-\log(x)\}$. Of these, setting $g_3(x) = \log(x)$ leads to the assumption $\theta(\mathbf{X}_i) = \exp(\gamma_0 + \boldsymbol{\gamma}^\top \mathbf{X}_i)$ of model (5), which motivates us to estimate the parameters $(\gamma_0, \boldsymbol{\gamma})$ based on pseudo-observations created by $\hat{\theta}_{\text{KM}}^i$ or $\hat{\theta}_{\text{NP}}^i$ via

$$U(\gamma_0, \boldsymbol{\gamma}) = \sum_{i=1}^n \frac{\partial g_3^{-1}(\gamma_0 + \boldsymbol{\gamma}^\top \mathbf{X}_i)}{\partial (\gamma_0, \boldsymbol{\gamma})} \mathbf{V}_{3,i}^{-1} \left(\hat{\theta}_i - g_3^{-1}(\gamma_0 + \boldsymbol{\gamma}^\top \mathbf{X}_i) \right) = 0, \tag{18}$$

where $\hat{\theta}_i$ is the pseudo-observations for θ , and $\mathbf{V}_{3,i}$ is a working variance of $\hat{\theta}_i$. Let $(\hat{\gamma}_{0,\text{KM}}^{\text{PO}}, \hat{\boldsymbol{\gamma}}_{\text{KM}}^{\text{PO}})$ and $(\hat{\gamma}_{0,\text{NP}}^{\text{PO}}, \hat{\boldsymbol{\gamma}}_{\text{NP}}^{\text{PO}})$ be the estimators obtained from (18) when $\hat{\theta}_i$ is replaced by $\hat{\theta}_{\text{KM}}^i$ and $\hat{\theta}_{\text{NP}}^i$ in (9) and (10), respectively. Under the assumption $\eta(\mathbf{Z}_i) = \exp(\boldsymbol{\phi}^\top \mathbf{Z}_i)$, we have $\log[-\log(1 - F^i(t))] = \boldsymbol{\phi}^\top \mathbf{Z}_i + \log[-\log(\bar{F}(t))]$, where $F^i(t) = 1 - \bar{F}(t)^{\eta(\mathbf{Z}_i)}$. We thus consider the GLM with

$$g_4(1 - \hat{F}^i(t_h)) = \varsigma_{t_h} + \boldsymbol{\phi}^\top \mathbf{Z}_i + \varepsilon_i, \tag{19}$$

where $\varsigma_{t_h} = \log\{-\log(\bar{F}(t_h))\}$, $t_h \in \mathbf{t} = \{t_1, \dots, t_H\}$, $g_4(x) = \log\{-\log(x)\}$, and $\varepsilon_i, i = 1, \dots, n$, are i.i.d. with mean zero. We consider the following GEE to estimate $\boldsymbol{\phi}$ and $\boldsymbol{\varsigma}_H$

$$U(\boldsymbol{\theta}) = \sum_{i=1}^n \frac{\partial \mathbf{g}_4^{-1}(\mathbf{t}, \boldsymbol{\theta}; \mathbf{Z}_i)}{\partial \boldsymbol{\theta}} \mathbf{V}_{4,i}^{-1} \left((1 - \hat{F}^i(\mathbf{t})) - \mathbf{g}_4^{-1}(\mathbf{t}, \boldsymbol{\theta}; \mathbf{Z}_i) \right) = 0, \tag{20}$$

where $\boldsymbol{\theta} = (\boldsymbol{\varsigma}_H, \boldsymbol{\phi})$, $\hat{F}^i(\mathbf{t}) = (\hat{F}^i(t_1), \dots, \hat{F}^i(t_H))^\top$ is the vector of pseudo-observations for $F^i(\mathbf{t})$ calculated at $\mathbf{t} = \{t_1, \dots, t_H\}$ for subject i , $\mathbf{g}_4^{-1}(\mathbf{t}, \boldsymbol{\theta}; \mathbf{Z}_i)$ is the H -column vector whose j^{th} component equals $\mathbf{g}_4^{-1}(\varsigma_{t_j} + \boldsymbol{\phi}^\top \mathbf{Z}_i)$, and $\mathbf{V}_{4,i}$ is a $H \times H$ working covariance matrix. Let $\hat{\boldsymbol{\theta}}_{\text{KM}}^{\text{PO}} = (\hat{\boldsymbol{\varsigma}}_{H,\text{KM}}^{\text{PO}}, \hat{\boldsymbol{\phi}}_{\text{KM}}^{\text{PO}})$ and $\hat{\boldsymbol{\theta}}_{\text{NP}}^{\text{PO}} = (\hat{\boldsymbol{\varsigma}}_{H,\text{NP}}^{\text{PO}}, \hat{\boldsymbol{\phi}}_{\text{NP}}^{\text{PO}})$ be the estimators obtained from equations (20) while $\hat{F}^i(\mathbf{t})$ is replaced by $\hat{F}_{\text{KM}}^i(\mathbf{t})$ and $\hat{F}_{\text{NP}}^i(\mathbf{t})$ in (11) and (12), respectively.

4.3 Variance estimation and model diagnosis

All estimators obtained from solving estimating equations mentioned in Sections 4.1 and 4.2 can be using the geese function in the R package *geepack*.³⁵ We adopt the approximate jackknife variance estimates,³⁶ which is available in the geese function. Note that adopting a sandwich estimator might lead to inconsistent and upward biased results for variance estimation; however, this has an insignificant impact in practical applications.²⁰ We follow the idea of pseudo-residuals³³ to assess the goodness-of-fit for (13) and (15) for the PHMC model. Define the pseudo-residuals $\{\hat{x}^i - g_1^{-1}(\hat{\alpha}_0 + \hat{\boldsymbol{\alpha}}^\top \mathbf{X}_i); i = 1, \dots, n\}$ based on either $\hat{\pi}_{\text{KM}}^i$ or $\hat{\pi}_{\text{NP}}^i$, and $\{\hat{S}_u^i(t) - g_2^{-1}(\hat{\xi}_t^{\text{PO}} + \hat{\boldsymbol{\beta}}^{\text{PO}\top} \mathbf{Z}_i); i = 1, \dots, n\}$ calculated at a given time $t \in \mathbf{t}$. If the model fits the data well, no trend should be perceptible when plotting residuals against a covariate. Similarly, we consider pseudo-residuals $\{\hat{\theta}_{\text{KM}}^i - g_3^{-1}(\hat{\gamma}_{0,\text{KM}}^{\text{PO}} + \hat{\boldsymbol{\gamma}}_{\text{KM}}^{\text{PO}\top} \mathbf{X}_i); i = 1, \dots, n\}$ or $\{\hat{\theta}_{\text{NP}}^i - g_3^{-1}(\hat{\gamma}_{0,\text{NP}}^{\text{PO}} + \hat{\boldsymbol{\gamma}}_{\text{NP}}^{\text{PO}\top} \mathbf{X}_i); i = 1, \dots, n\}$ for model (17) and the pseudo-residuals $\{(1 - \hat{F}_{\text{KM}}^i(\mathbf{t})) - g_4^{-1}(\boldsymbol{\varsigma}_{\mathbf{t},\text{KM}}^{\text{PO}} + \hat{\boldsymbol{\phi}}_{\text{KM}}^{\text{PO}\top} \mathbf{Z}_i); i = 1, \dots, n\}$ or $\{(1 - \hat{F}_{\text{NP}}^i(\mathbf{t})) - g_4^{-1}(\boldsymbol{\varsigma}_{\mathbf{t},\text{NP}}^{\text{PO}} + \hat{\boldsymbol{\phi}}_{\text{NP}}^{\text{PO}\top} \mathbf{Z}_i); i = 1, \dots, n\}$ at a given time $t \in \mathbf{t}$ for model (19). The idea of pseudo-residuals is illustrated in the melanoma data in Section 6.1.

4.4 Variable selection

The proposed pseudo-observations approach allows variable selection and parameter estimation to be simultaneously implemented in each component of the PHMC model and the PHPH model by penalizing the corresponding GEEs.²⁸ Specifically, for the *incidence* component of the PHMC model, we penalize the GEE in (14) by

$$S(\alpha_0, \boldsymbol{\alpha}) = U(\alpha_0, \boldsymbol{\alpha}) - \mathbf{q}_{\lambda_1}(|\boldsymbol{\alpha}|) \circ \text{sign}(\boldsymbol{\alpha}), \quad (21)$$

where for some $p \times 1$ vectors \mathbf{u} and \mathbf{v} , $\mathbf{q}_{\lambda_1}(\mathbf{u}) = \{q_{\lambda_1}(|u_1|), \dots, q_{\lambda_1}(|u_p|)\}^\top$ is a vector of penalty functions for some tuning parameter λ_1 , $\text{sign}(\mathbf{u}) = \{\text{sign}(u_1), \dots, \text{sign}(u_p)\}^\top$, $\text{sign}(x) = I(x > 0) - I(x < 0)$, and $\mathbf{u} \circ \mathbf{v}$ is the element-wise product of \mathbf{u} and \mathbf{v} . The intercept term α_0 is not penalized and has been left out of the penalty term in (21). Similarly, we consider the following PGEE for $\boldsymbol{\psi}$ for the *latency* component of the PHMC model by extending (16)

$$S(\boldsymbol{\psi}) = U(\boldsymbol{\psi}) - \mathbf{q}_{\lambda_2}(|\boldsymbol{\beta}|) \circ \text{sign}(\boldsymbol{\beta}), \quad (22)$$

where λ_2 is the tuning parameter, and the intercept term, $\boldsymbol{\xi}_H$, is left out of the penalty term.

For the *long-term* effect of the PHPH model, we extend (18) to the following PGEE

$$S(\gamma_0, \boldsymbol{\gamma}) = U(\gamma_0, \boldsymbol{\gamma}) - \mathbf{q}_{\lambda_3}(|\boldsymbol{\gamma}|) \circ \text{sign}(\boldsymbol{\gamma}), \quad (23)$$

where λ_3 is the tuning parameter and the intercept term γ_0 is not penalized.

Lastly, for the *short-term* effect of the PHPH model, we penalize (20) through

$$S(\boldsymbol{\Theta}) = U(\boldsymbol{\Theta}) - \mathbf{q}_{\lambda_4}(|\boldsymbol{\phi}|) \circ \text{sign}(\boldsymbol{\phi}), \quad (24)$$

where λ_4 is the tuning parameter and $\boldsymbol{\varsigma}_{t_H}$ is not penalized. In the simulation studies and data analyzes, we illustrate the proposed procedure with the SCAD penalty²⁴ and select the tuning parameters via five-fold cross-validation, where the data are randomly partitioned into five subsets of approximately equal sizes. For a given tuning parameter λ , we calculate the overall cross-validated prediction error $CV(\lambda) = |N^{(k)}|^{-1} \sum_{j \in N^{(k)}} m_j^{-1} \sum_{\ell=1}^{m_j} (PR(\hat{\boldsymbol{\theta}}^{(-k)\top} \mathbf{W}_{j\ell}))^2$ where $\hat{\boldsymbol{\theta}}^{(-k)}$ is the PGEE estimator based on data without the k th subset, $|N^{(k)}|$ is the size of the k th subset, m_j is the number of pseudo-observations for subject j , and $PR(\hat{\boldsymbol{\theta}}^{(-k)\top} \mathbf{W}_{j\ell})$ are the pseudo-residuals as defined in Section 4.3 with covariates $\mathbf{W}_{j\ell}$. We use the estimates obtained from the unpenalized GEEs as the initial values in the iteration of PGEEs and cross-validation.

5 Simulation

Simulation studies are conducted to assess the finite sample performance of proposed estimators. We first evaluate the performance under the PHMC model. Specifically, we generate the cure status according to the logistic model (2) with one covariate X_i , $i = 1, \dots, n$, generated from a Bernoulli(0.5) distribution. The regression coefficient is set at $\boldsymbol{\alpha} = (\alpha_0, -1)$ so that the treatment group (e.g. those with $X_i = 1$) are more likely to be cured. The intercept is chosen to be $\alpha_0 = 2.8$, $\alpha_0 = 2$ or $\alpha_0 = 0.9$ to achieve the average cure rates of 10% (14.1% among those with $X_i = 1$), 20% (26.7% among those with $X_i = 1$) or 40% (52.3% among those with $X_i = 1$), respectively. On the other hand, the survival times are generated from the Cox PH model (3), with $\lambda_0(t) = 1/3$, and (Z_{i1}, Z_{i2}) generated from independent Bernoulli(0.5) and Uniform(0, 1), respectively. We set the regression coefficient in (3) at $\boldsymbol{\beta} = (1, 0.5)$ and generated the censoring times from Uniform(0, c), where c is chosen to make 10% of the censored subjects susceptible. Throughout the simulations, the pseudo-observations are calculated at 10 time-points from the quantiles of observed event times between 0 and t_{\max} .

Table 1 summarizes the simulation results of 20% and 40% cure rates scenarios based on (14) and (16) with link functions $g_1(x) = \log\{x/(1-x)\}$, $g_2(x) = \log\{-\log(x)\}$ and $\xi_{t_h} = \log \Lambda_0(t_h)$. The scenario with 10% cure rate is presented in Table 1 in Web Appendix E. We only present the results with a working independence assumption among the pseudo-observations. Adopting a more complicated covariance structure provides no obvious improvement as presented in Klein and Andersen³⁷ and Graw et al.¹⁹ The proposed estimates are compared with the EM-algorithm estimators obtained from Peng and Dear⁶ with $B = 100$ bootstrap samples for standard error estimation. For each estimator, we report the average bias (Bias), the empirical standard error (ESE), the average of the standard error estimator (SEE), and the empirical coverage rate (CR) of 95% confidence interval based on 500 replicates with sample size $n = 200, 400, 600$, and 1000. Overall, the proposed estimators perform reasonably in all scenarios with Bias, ESE, and SEE all decreasing with increasing n while CRs are close to the nominal level of 0.95. The estimators $(\hat{\boldsymbol{\alpha}}_{0, \text{NP}}^{\text{PO}}, \hat{\boldsymbol{\alpha}}_{\text{NP}}^{\text{PO}})$ and $(\hat{\boldsymbol{\alpha}}_{0, \text{KM}}^{\text{PO}}, \hat{\boldsymbol{\alpha}}_{\text{KM}}^{\text{PO}})$ have similar performance, indicating that pseudo-observations constructed by (6) or (7) are both appropriate. Compared to the estimates of Peng and Dear,⁶ our

Table 1. Simulation summaries under the PHMC model based on 500 replicates.

n		Incidence					Latency				
		$\hat{\alpha}_{0,NP}^{PO}$	$\hat{\alpha}_{1,NP}^{PO}$	$\hat{\alpha}_{0,KM}^{PO}$	$\hat{\alpha}_{1,KM}^{PO}$	$\hat{\alpha}_0^{EM}$	$\hat{\alpha}_1^{EM}$	$\hat{\beta}_1^{PO}$	$\hat{\beta}_2^{PO}$	$\hat{\beta}_1^{EM}$	$\hat{\beta}_2^{EM}$
20% cure rate, 30% censoring rate											
200	Bias	0.040	-0.037	0.038	-0.034	0.048	-0.026	0.011	0.005	0.009	0.001
	ESE	0.431	0.507	0.423	0.503	0.396	0.468	0.226	0.379	0.209	0.321
	SEE	0.473	0.560	0.469	0.556	0.349	0.477	0.232	0.381	0.207	0.335
	CR	0.954	0.963	0.957	0.963	0.923	0.968	0.964	0.966	0.960	0.968
400	Bias	0.018	-0.027	0.019	-0.028	0.032	-0.036	0.016	0.022	0.013	0.017
	ESE	0.328	0.377	0.327	0.377	0.286	0.325	0.159	0.277	0.138	0.230
	SEE	0.323	0.387	0.324	0.387	0.271	0.351	0.166	0.272	0.142	0.232
	CR	0.956	0.962	0.956	0.960	0.930	0.958	0.960	0.964	0.958	0.960
600	Bias	0.014	-0.013	0.011	-0.012	0.021	-0.017	0.009	-0.013	0.007	-0.013
	ESE	0.260	0.300	0.254	0.299	0.215	0.251	0.124	0.233	0.109	0.190
	SEE	0.266	0.316	0.264	0.315	0.205	0.272	0.136	0.223	0.115	0.188
	CR	0.960	0.954	0.964	0.950	0.956	0.976	0.970	0.932	0.974	0.936
1000	Bias	-0.010	0.008	0.008	0.010	0.015	-0.014	-0.001	-0.004	-0.002	0.000
	ESE	0.216	0.236	0.217	0.246	0.164	0.188	0.093	0.172	0.086	0.143
	SEE	0.201	0.230	0.199	0.238	0.161	0.206	0.104	0.173	0.088	0.144
	CR	0.940	0.954	0.954	0.958	0.940	0.974	0.964	0.946	0.958	0.946
40% cure rate, 50% censoring rate											
200	Bias	-0.027	-0.001	-0.025	-0.001	-0.012	-0.008	0.011	-0.045	-0.008	-0.044
	ESE	0.334	0.425	0.335	0.425	0.288	0.363	0.311	0.586	0.279	0.404
	SEE	0.312	0.411	0.312	0.411	0.242	0.368	0.328	0.556	0.271	0.439
	CR	0.926	0.958	0.926	0.958	0.886	0.950	0.954	0.946	0.936	0.964
400	Bias	-0.018	-0.019	-0.019	-0.020	-0.021	-0.004	-0.010	-0.014	-0.019	-0.010
	ESE	0.290	0.339	0.298	0.351	0.192	0.234	0.228	0.492	0.192	0.299
	SEE	0.266	0.331	0.283	0.347	0.162	0.251	0.239	0.412	0.184	0.298
	CR	0.938	0.974	0.934	0.976	0.892	0.966	0.944	0.942	0.944	0.954
600	Bias	-0.018	0.015	-0.018	0.015	-0.018	0.007	-0.012	-0.032	-0.008	-0.008
	ESE	0.169	0.224	0.169	0.225	0.156	0.199	0.161	0.367	0.137	0.248
	SEE	0.173	0.227	0.172	0.227	0.153	0.202	0.185	0.316	0.147	0.237
	CR	0.940	0.956	0.940	0.960	0.940	0.950	0.965	0.945	0.962	0.932
1000	Bias	-0.015	0.016	-0.015	0.009	-0.010	0.002	-0.024	-0.023	-0.017	-0.010
	ESE	0.129	0.168	0.128	0.168	0.119	0.153	0.119	0.234	0.108	0.185
	SEE	0.128	0.169	0.129	0.170	0.120	0.154	0.134	0.229	0.111	0.182
	CR	0.942	0.945	0.942	0.952	0.940	0.956	0.958	0.938	0.940	0.942

Bias: bias of parameter estimator; ESE: the empirical standard error; SEE: average of the standard error estimator; CR: coverage rate of the 95% confidence interval; PHMC: proportional hazards mixture cure; $(\hat{\alpha}_0^{EM}, \hat{\alpha}_1^{EM})$ and $(\hat{\beta}_1^{EM}, \hat{\beta}_2^{EM})$ are EM-algorithm based estimators with standard errors are estimated based on $B = 100$ bootstrap samples⁶ which can be implemented via the R package *smcure*.³⁸

estimators have higher ESE and SEE under various sample sizes, indicating a loss of efficiency. Specifically, based on our simulation settings, the smallest relative efficiency loss for the *latency* component is around 12% with $n = 1000$ under the 10% cure rate scenario. The smallest relative efficiency loss for the *incidence* component is around 9% with $n = 1000$ under the 40% cure rate scenario. This loss efficiency situation is likely due to a discrete approximation to the baseline hazard function with a continuous time scale, for example, 10 selected time-points for the latency component in our approach. Note that the latency component yields smaller ESE and SEE under 10% cure rate compared to 20% and 40% cure rates scenarios as more noncured subjects contribute to the estimation of β .

On the other hand, Table 2 in Web Appendix E reports the average computing time in seconds for each estimator under the 10% cure rate scenario based on 500 replicates. For each fixed n , our proposed estimators including variance estimation have smaller total computing times (summation of latency and incidence) than that of the EM-algorithm estimate⁶ using $B = 100$ bootstrap samples for standard error estimation. Computing times increase as long as the sample size increases. Among our proposed estimators for the incidence component, the computing time for the KM estimator (6) is faster than that of the estimator (7). Note that all results of computing times are implemented in R and performed on a Linux machine with an Intel Core i7-8565U processor and 15.4 GB memory.

We next evaluate the performance of the proposed estimators under the PHPH model (5). For subject i , two independent covariates X_{i1} and X_{i2} are generated from Bernoulli(0.5) and standard normal distribution, respectively. We set $\theta(\mathbf{X}_i) = \exp(\gamma_0 + \gamma_1 X_{i1})$, $\eta(\mathbf{Z}_i) = \exp(\phi_1 X_{i1} + \phi_2 X_{i2})$ and $\bar{F}(x) = \exp(-2x)$. It is not straightforward to simulate survival time from the improper survival function of model (5) as it has a positive mass at ∞ . We thus utilize the connection between the MC model (1) and the PHPH model (5) to generate survival times. The data generation algorithm is summarized in Web Appendix D. We set $\gamma_1 = -0.1$, $(\phi_1, \phi_2) = (0.4, -0.3)$ with two different values of γ_0 , 0.5 and -0.05 , which correspond to 20% and 40% cure rate, respectively. The censoring times are independently generated from Uniform(0, c), where c is chosen so that 10% of the censored subjects are susceptible. Tables 3 and 4 presented in Web Appendix E show the results of estimators $(\hat{\gamma}_{0,KM}^{PO}, \hat{\gamma}_{1,KM}^{PO})$ and $(\hat{\gamma}_{0,NP}^{PO}, \hat{\gamma}_{1,NP}^{PO})$ obtained from (18) and estimators $(\hat{\phi}_{1,NP}^{PO}, \hat{\phi}_{2,NP}^{PO})$ and $(\hat{\phi}_{1,KM}^{PO}, \hat{\phi}_{2,KM}^{PO})$ obtained from (20), respectively. A maximum likelihood estimator (MLE) is implemented for comparison. The proposed estimates are virtually unbiased. The SEE is reasonably close to the ESE, and the CR is close to the nominal level. The two constructions of the pseudo-observations yield a similar pattern, indicating that both approaches are feasible for constructing pseudo-observations. The proposed estimators have higher ESE and SEE than MLE under different sample sizes, indicating a loss of efficiency. For example, the relative efficiency losses are around 19% and 25% for *long-term* and *short-term* effects, respectively, under the scenario, $n = 1000$, 40% cure rate, and 50% censoring rate.

To study the performance of variable selection under the PHMC model, we consider the covariate $\mathbf{X} = (X_1, \dots, X_{20})$ in which X_1, X_2 are independently generated from Uniform(0,1), X_3 and X_4 are independently generated from Bernoulli(0.5), and X_5, \dots, X_{20} are generated from a multivariate normal distribution with $E(X_i) = 0$ and $Cov(X_i, X_j) = 0.5^{|i-j|}$, for $i, j = 5, \dots, 20$. We set the regression coefficients at $\alpha_0 = 1.1$, $\boldsymbol{\alpha} = (0, 1, -1.2, 0, 0, -0.9, 0.8, 0, 0, \dots, 0)^T$, and $\boldsymbol{\beta} = (-0.7, 0, 1, 0, -0.5, 0.8, 0, 0, 0, \dots, 0)^T$, and the baseline hazard function $\lambda_0(t) = 1/3$. Those configurations yield a cure rate of 30%. The censoring times are independently generated from Uniform(0, c), where c is chosen so that either 10% or 30% of the censored subjects are susceptible. For each simulated data set, we fit the GEE full model that considers all covariates, the GEE Oracle model that only includes covariates with nonzero coefficients, the proposed PGEE model with SCAD penalty, and the PHMC model with LASSO and ALASSO penalties proposed by Masud et al.²⁷

Table 2 summarizes the results for variable selection under the PHMC model with $n = 200$ and 600. Based on 200 replicates, the average mean square error (MSE), true positives (TP), and false positives (FP) are reported. The MSE is defined as $\sum_{j=1}^{200} \|\hat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}\|^2 / 200$ for *incidence* component and $\sum_{j=1}^{200} \|\hat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}\|^2 / 200$ for *latency* component, where $\hat{\boldsymbol{\alpha}}^j$ and $\hat{\boldsymbol{\beta}}^j$ are the estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ based on the j th generated dataset. The TP and FP are calculated as the average number of selected covariates that have actual nonzero and zero coefficients, respectively. Five-fold cross-validation is used to determine the tuning parameter. For the *incidence* component, the two pseudo-observations approaches yield similar performances, making them practically identical. For all scenarios, the MSE of proposed PGEE approaches is smaller than that of the full model but is larger than that of the oracle model. However, it becomes closer to the MSE of the oracle model as n increases. In addition, when the censoring rate is 40%, the proposed PGEE approaches behave closer to the oracle model, have TP closer to the number of nonzero covariates, and decreasing FP as n increases. On the contrary, the FP of the LASSO and ALASSO estimators do not seem to decrease with increasing n . For the *latency* component, we consider the proposed PGEE with three different correlation structures: independence, exchangeable, and AR(1). The results of PGEE with independent structure tend to have slightly higher TP and FP and lower MSE than that of the PGEE with exchangeable and AR(1) correlation structure. However, the performances are close when the sample size is large and with 40% censoring rate, indicating that little advantage gains while specifying complicated correlation structure among pseudo-observations for variable selection. When the censoring rate is 60%, the TP decreases and the MSE increases for all the presented methods. Compared to the results based on LASSO and ALASSO, our proposed PGEE performs reasonably in identifying important variables with $n = 600$. The results with $n = 400$ and 1000 reveal a similar trend and are presented in Table 5 in Web Appendix E.

To investigate variable selection performance under the PHPH model, we generate the covariate vector following the PHMC model's configuration to specify the *short-term* and *long-term* effects. In this case, we set $\gamma_0 = 0.85$, $\boldsymbol{\gamma} = (0, 0, -0.9, 0, 0, -0.7, 0, 1, 0, \dots, 0)$, $\boldsymbol{\phi} = (-0.5, 0, 0.8, 0, -0.7, 0, 0, 0, 0, \dots, 0)$, and the baseline function $\bar{F}(x) = \exp(-2x)$, resulting in a cure rate of 30%. The censoring times are independently generated from Uniform(0, c), where c is chosen so that either 10% or 30% of the censored subjects are susceptible. Tables 6 and 7 in Web Appendix E depict the simulation results for *long-term* and *short-term* effects based on different sample sizes. The definition of the MSE for the *long-term* and *short-term* effects is similar to that of the *incidence* and *latency* components mentioned in the above paragraph. We observe that the two constructions of pseudo-observations yield similar results. For each sample size n , the MSE of the PGEE is smaller than that of the full model but is larger than that of the oracle model. As n increases, the MSE of the PGEE decreases and is close to that of the Oracle model. Also, the TP increases and the FP decreases. However, as expected, the MSE increases and the TP decreases when the censoring rate increases to 60%. Finally, based on our simulation results, we observe minimal performance gains when specifying complicated correlations among

Table 2. Simulation summaries for variable selection on the *incidence* and *latency* component of the PHMC model with 30% cure rate.

n		40% censoring rate			60% censoring rate		
		MSE	TP	FP	MSE	TP	FP
Incidence							
200	Full.NP	7.44	–	–	10.96	–	–
	Full.KM	7.45	–	–	10.96	–	–
	Oracle.NP	1.66	4	0	4.54	4	0
	Oracle.KM	1.65	4	0	4.54	4	0
	SCAD.NP	5.23	2.77	7.40	6.21	1.49	4.87
	SCAD.KM	5.18	2.77	7.59	6.36	1.49	4.72
	LASSO	3.33	1.12	0.78	4.56	0.26	0.15
	ALASSO	3.23	1.26	0.69	4.54	0.23	0.19
	600	Full.NP	1.39	–	–	4.96	–
Full.KM		1.35	–	–	4.99	–	–
Oracle.NP		0.39	4	0	3.57	4	0
Oracle.KM		0.38	4	0	3.63	4	0
SCAD.NP		0.68	3.87	0.42	4.64	0.98	0.63
SCAD.KM		0.68	3.86	0.42	4.69	1.00	0.70
LASSO		1.20	3.81	2.06	4.24	0.86	0.86
ALASSO		1.17	3.80	1.88	4.23	0.93	0.66
Latency							
200	Full.indep	1.61	–	–	3.87	–	–
	Full.exch	2.39	–	–	4.44	–	–
	Full.ar I	2.23	–	–	4.78	–	–
	Oracle.indep	0.42	4	0	1.06	4	0
	Oracle.exch	0.41	4	0	1.80	4	0
	Oracle.ar I	0.41	4	0	1.30	4	0
	SCAD.indep	0.99	3.17	2.32	2.40	0.54	0.95
	SCAD.exch	1.08	2.86	0.75	2.40	0.31	0.35
	SCAD.ar I	1.04	2.98	0.86	2.51	0.28	0.20
	LASSO	1.38	2.26	1.70	2.26	0.34	0.36
	ALASSO	0.67	3.34	0.79	1.90	1.36	0.58
600	Full.indep	0.35	–	–	1.30	–	–
	Full.exch	0.41	–	–	2.17	–	–
	Full.ar I	0.37	–	–	1.36	–	–
	Oracle.indep	0.12	4	0	0.58	4	0
	Oracle.exch	0.13	4	0	0.71	4	0
	Oracle.ar I	0.12	4	0	0.61	4	0
	SCAD.indep	0.21	3.87	0.77	1.20	2.38	1.50
	SCAD.exch	0.24	3.82	0.18	1.54	1.89	0.86
	SCAD.ar I	0.24	3.83	0.16	1.23	1.72	0.38
	LASSO	0.31	3.94	2.64	1.48	2.35	1.43
	ALASSO	0.14	3.96	0.27	0.66	3.45	0.49

Full: model includes all covariates; Oracle: model only includes the covariates with nonzero coefficients; Acronyms that ends with .NP or .KM indicates pseudo-observations $\hat{\beta}_{NP}^j$ or $\hat{\beta}_{KM}^j$, respectively; Acronyms that ends with .indep, .exch and .ar I indicates independence, exchangeable and AR(1) correlation structure among pseudo-observations, respectively; LASSO: PHMC model with LASSO penalty; ALASSO: PHMC model with ALASSO penalty; MSE: the average estimated mean square error; TP: the average true positives; FP: the average false positives; PHMC: proportional hazards mixture cure.

pseudo-observations. With the same simulation specifications, we observe an improvement in FP and a loss in TP (results not shown here) when a one-standard error rule is used to select the optimum tuning parameter.

6 Data analysis

6.1 The melanoma data

We apply the proposed method to a melanoma dataset from the Eastern Cooperative Oncology Group phase III clinical trial e1684,¹ which is available from the R package *smcure*.³⁸ The primary objective is to determine whether the high dose

interferon alpha-2b (IFN) regimen in postoperative adjuvant therapy would lead to a significantly prolonged interval of relapse-free for melanoma. The event of interest is the relapse of melanoma. The interested covariates include treatment (0=placebo, 1=IFN), gender (0=male, 1=female), and age (centered to zero). After excluding missing data, a total of 284 subjects is included in the analysis. The overall censoring rate is 30.9%. Figure 1(a) shows the KM estimates stratified by treatment and gender. The KM estimates level off at the end of the study, suggesting a fraction of nonsusceptibility to the recurrence of melanoma. This observation is confirmed by the Maller-Zhou test³⁰ with a p -value of <0.001 . Based on the KM curves, a male has a higher cure rate than a female in the treatment group, whereas it is reversed in the control group, implying an interaction between the treatment and gender. Thus, the interaction term is also considered in the data analysis.

The top panel of Table 3 presents the results from the PHMC model obtained by (14) and (16) with $g_1(x) = \log \{x/(1-x)\}$, $g_2(x) = \log \{-\log(x)\}$, and $\xi_{t_h} = \log \Lambda_0(t_h)$. The lower panel of Table 3 presents the results from the PHPH model obtained by (18) and (20) with $g_3(x) = \log(x)$, $g_4(x) = \log \{-\log(x)\}$, and $\zeta_{t_h} = \log \{-\log(\bar{F}(t_h))\}$. For comparison, we included the estimator⁶ based on the EM-algorithm with standard errors obtained from 500 bootstrapped samples. The two constructions $\hat{\pi}_{KM}^i$ and $\hat{\pi}_{NP}^i$ of the pseudo-observations for π yield similar patterns. For the incidence component, our proposed estimates $\hat{\alpha}_{NP}^{PO}$ and $\hat{\alpha}_{KM}^{PO}$ are similar to the estimate $\hat{\alpha}^{EM}$. The treatment has a significantly adverse effect on the susceptibility (noncured), which means the treatment substantially improves the relapse-free

Table 3. Parameter estimates for the melanoma data.

PHMC model									
Incidence	$\hat{\alpha}^{EM}$			$\hat{\alpha}_{NP}^{PO}$			$\hat{\alpha}_{KM}^{PO}$		
	Est.	SEE	p -value	Est.	SEE	p -value	Est.	SEE	p -value
Intercept	1.502	0.366	<0.001	1.671	0.563	0.003	1.737	0.584	0.003
Treatment	-0.866	0.426	0.042	-1.272	0.612	0.037	-1.294	0.633	0.040
Gender	-0.442	0.492	0.368	-0.614	0.614	0.315	-0.628	0.635	0.320
Age	0.018	0.014	0.189	0.024	0.013	0.066	0.024	0.013	0.066
Treatment:Gender	0.648	0.665	0.329	0.886	0.734	0.225	0.903	0.755	0.228
Latency									
	$\hat{\beta}^{EM}$			$\hat{\beta}^{PO}$					
	Est.	SEE	p -value	Est.	SEE	p -value	Est.	SEE	p -value
Treatment	-0.012	0.236	0.958	-0.009	0.299	0.975			
Gender	0.266	0.258	0.303	0.151	0.310	0.629			
Age	-0.007	0.007	0.266	-0.009	0.007	0.206			
Treatment: Gender	-0.334	0.369	0.365	-0.151	0.383	0.696			
PHPH model									
Long-term effect	$\hat{\gamma}_{NP}^{PO}$			$\hat{\gamma}_{KM}^{PO}$					
	Est.	SEE	p -value	Est.	SEE	p -value	Est.	SEE	p -value
Intercept	0.519	0.167	0.002	0.538	0.164	0.001			
Treatment	-0.689	0.275	0.012	-0.668	0.266	0.012			
Gender	-0.266	0.223	0.229	-0.259	0.218	0.231			
Age	0.012	0.006	0.044	0.012	0.006	0.044			
Treatment: Gender	0.467	0.390	0.228	0.453	0.378	0.227			
Short-term effect	$\hat{\phi}_{NP}^{PO}$			$\hat{\phi}_{KM}^{PO}$					
	Est.	SEE	p -value	Est.	SEE	p -value	Est.	SEE	p -value
Treatment	0.586	0.500	0.247	0.565	0.487	0.252			
Gender	0.557	0.482	0.253	0.537	0.469	0.258			
Age	-0.017	0.010	0.104	-0.017	0.010	0.101			
Treatment: Gender	-0.718	0.572	0.214	-0.689	0.557	0.220			

Est.: Parameter estimate; SEE: Standard error estimate; $\hat{\alpha}^{EM}$ and $\hat{\beta}^{EM}$ are EM-algorithm estimators with standard errors are estimated based on 500 bootstrap samples obtained from the R package smcure.³⁸

(cured) rate, especially for males. The positive age effect indicates that older patients tend to have a higher relapse rate of melanoma. In the *latency* component, none of the four covariates are significantly associated with the failure time if patients are susceptible. However, female patients tend to have a lower risk of recurring melanoma than male patients in the treatment group but higher in the control group.

Under the PHPH model, the treatment has a significantly negative effect on the *long-term* effect, indicating that high dose IFN regimen increase the rate of relapse-free melanoma, especially for males. Older patients have a high possibility of recurring melanoma even though the effect is not statistically significant at the 5% nominal level. Those results are consistent with the findings in the *incidence* of the PHMC model. None of the four covariates are statistically significant at the 5% nominal level related to the *short-term* effect. The estimates indicate females are likely to have a more rapidly developing melanoma within the treatment group. However, the result is reversed in the control group, i.e., males are expected to experience melanoma sooner than females. Those findings are verified in Figure 1(a), where the KM estimate drops relatively faster for the females than for the males in the treatment group. In contrast, the KM estimates drop faster for the males than females in the control group.

To understand the covariate effect of cure rates in two models, we calculate the estimated cure rates in each group while holding the age variable at the average. The results presented in Table 6 of Web Appendix E suggests that there are higher cure rates among the treatment groups, echoing our finding that the treatment substantially improves the relapse-free (cured) rate. The estimated cure rates based on $\hat{\alpha}_{NP}^{PO}$ and $\hat{\alpha}_{KM}^{PO}$ under the PHMC model are close to that based on $\hat{S}_{KM}(t_{max})$; however, the results based on $\hat{\gamma}_{NP}^{PO}$ and $\hat{\gamma}_{KM}^{PO}$ under the PHPH model tend to have higher cure rate except for the female in the treatment group. This suggests that the PHMC model is more conservative in estimating cure rates. To assess the goodness of fit, Figure 2 in Web Appendix F presents the boxplots of pseudo-residuals for models (13) and (15) of the PHMC model. The top panel shows the boxplots of pseudo-residuals stratified by treatment and gender based on the pseudo-observations $\hat{\pi}_{KM}^i$ from (6) and its corresponding estimators. The pseudo-residuals fluctuate around zero, indicating the adequacy of the proposed GLM even though the pseudo-residuals have a larger variation in the Control/Male group. The bottom panel illustrates the boxplots of pseudo-residuals based on pseudo-observations $\hat{S}_u^i(t)$ calculated at four given time points chosen from the quantiles of observed event times. The residuals are symmetric around 0 at any given time point, which implies the proposed GLM model fits the data well even though the variation of the pseudo-observations increases in the Control/Male group as the time increases. The boxplots of pseudo-residuals stratified by treatment and gender based on GLMs (17) and (19) under the PHPH model are also presented in Figure 3 of Web Appendix F. The top panel shows the residuals calculated based on pseudo-observations $\hat{\theta}_{KM}^i$ and its resulting estimators, and the bottom panel presents the pseudo-residuals calculated based on pseudo-observations $\hat{F}_{KM}^i(t)$ at four given time points. Compared to Figure 2 in Web Appendix F, the pseudo-residuals under the PHPH model tend to have more considerable variations. This result might suggest that the PHMC model fits the data better than the PHPH model.

6.2 The dental data

We apply the proposed PGEE approaches to a dental dataset from the Creighton University School of Dentistry.² The dataset contains dental records from 5336 patients with periodontal disease collected between August 2007 and March 2013. In this work, the outcome of interest is the time to the first tooth loss due to periodontal reasons for each patient, yielding a censoring rate of 74.1%. The data analysis includes a total of 44 risk factors, whose detailed descriptions can be found in Tables 3 and 4 in Calhoun et al.² The length of the follow-up was 5.7 years, and the last event occurred at 5.37 years for both molar and non-molar groups. There were 35 and 20 teeth censored between the last event and the end of the study for the molar group and non-molars group, respectively. Figure 1(b) shows the KM survival curves stratified by the tooth type (3456 molars vs. 1880 non-molars) leveling off to nonzero probabilities, indicating a possible presence of a cured fraction in the population. This is also confirmed by the Maller-Zhou test³⁰ with a p -value of <0.001 .

As the study aims to identify which factors are associated with tooth loss, the proposed PGEE provides a logical tool for risk factor screening. We perform the variable selection procedure in both the PHMC model and the PHPH model. Under the PHMC model, we apply the proposed PGEEs in (21) and (22) with pseudo-observations created by $\hat{\pi}_{KM}^i$ and $\hat{S}_u^i(t)$, respectively. While under the PHPH model, we apply the proposed PGEEs in (23) and (24) with pseudo-observations created by $\hat{\theta}_{KM}^i$ and $\hat{F}_{KM}^i(t)$, respectively. In both the PHMC model and the PHPH model, a working independence correlation is incorporated for the ten pseudo-observations obtained from the quantiles of observed event times. The tuning parameters are selected via five-fold cross-validation. For comparison, we also fit the PHMC model with LASSO and ALASSO penalties proposed by Masud et al.²⁷ implemented in the R package *intsurv*.³⁹ In the variable selection procedure, the molar variable's coefficient is not penalized since the variable effect is the research of interest. Tables 4 and 5 present the variable selection results. Under the PHMC assumption, when predicting whether a tooth is cured, the LASSO and ALASSO selected more variables as we observed in the simulation studies; see Table 2 and Table 5 in Web Appendix E. The

Table 4. Estimated coefficients based on different penalization approaches for the dental dataset.

	PHMC						PHPH	
	LASSO		ALASSO		PGEE		PGEE	
	Incidence	Latency	Incidence	Latency	Incidence	Latency	Long-term	Short-term
Tooth-level factors								
molar	-0.327	-0.268	-0.392	-0.224	-0.618	-0.063	-1.303	-0.048
mobil	0.541	0.328	0.718		1.125	0.751		0.671
bleed	0.003		0.002		0.002			-0.003
plaque						0.002	0.013	0.002
pocket_mean	0.428		0.417		0.746			
pocket_max								
cal_mean								
cal_max	0.063		0.066					
fgm_mean								
fgm_max								
filled					1.372			
decay_new	0.063	0.111	0.119		1.591	0.815		0.744
decay_recurrent	0.194		0.165		1.619	0.590		0.547
dfs					-1.295			
crown	0.020							
endo	1.091		1.088		2.074			-0.683
filled tooth	-0.431	-0.066	-0.450		-0.690	-0.859		-0.842
decayed tooth	0.524		0.577			-1.434		-1.522

Table 5. (Continuation of Table 4) Estimated coefficients based on different penalization approaches for the dental dataset.

	PHMC						PHPH	
	LASSO		ALASSO		PGEE		PGEE	
	Incidence	Latency	Incidence	Latency	Incidence	Latency	Long-term	Short-term
Subject-level factors								
bleed_ave						0.006	-0.019	0.006
plaque_ave						-0.007	0.041	-0.007
pocket_mean_ave	0.013		0.021					
pocket_max_ave	0.017		0.023					
cal_mean_ave								
cal_max_ave								
fgm_mean_ave								
fgm_max_ave								
filled_sum						-0.007	0.011	-0.001
filled_ave								
decay_new_sum			0.016			0.005		
decay_new_ave	0.788			0.493	1.286			
decay_recur_sum	0.097		0.087		0.427	-0.009	0.015	-0.028
decayed_recur_ave					-3.686			
dfs_sum						-0.002	0.033	-0.006
dfs_ave								
filled_tooth_sum						0.033		0.033
filled_tooth_ave								
decayed_tooth_sum	0.044		0.049				0.034	
decayed_tooth_ave					-0.981	0.619		0.641
missing_tooth_sum								
missing_tooth_ave					-1.683	0.597		0.988
total_tooth	0.002							
Demographic factors								
age at baseline	0.013		0.013			0.005	0.009	0.005
gender	-0.051		-0.032					
Health factors								
diabetes						0.589		0.654
Tobacco use	0.171		0.159					0.362

common selected factors based on presented methods include mobility score (mobil), bleeding on probing (bleed), periodontal probing depth (pocket_mean), decayed surfaces new (decayed_new), decayed surfaces recurrent (decayed_recurrent), endodontic therapy (endo), filled tooth (filled_tooth) and recurrent decayed surfaces (decay_recur_sum). Only the variable filled tooth increases the chance of cure while others decrease the chance of cure. In addition, for the *latency*, the PGEE tends to identify more risk factors in predicting tooth survival time. The mobility score (mobil), decayed surfaces new (decayed_new) and filled tooth (filled_tooth) are three common selected factors related to the tooth survival time based on PGEE and LASSO. Under the PHPH model assumption, our proposed PGEE approach suggests that factors plaque score (plaque), bleeding on probing (bleed_ave), plaque index (plaque_ave), filled surfaces (filled_sum), recurrent decayed surfaces (decay_recur_sum), decayed and filled surfaces (dfs_sum), number of decayed teeth (decayed_tooth_sum) and age at baseline (baseline_age) are importantly associated with the *long-term* effect. Moreover, 20 factors are identified in association with the *short-term* effect, indicating losing the tooth more rapidly. Interestingly, we observe that 15 variables are both selected by the PGEE in the *latency* component of the PHMC model and the *short-term* effect of the PHPH model even though covariates are interpreted differently in the two models.

7 Conclusions

This article extends the pseudo-observations approach to the context of right-censored survival data with a cure fraction for two popular cure models, the MC model and the BCH model. Several estimators for the regression parameters related to the cure rate and the risk of experiencing the event are proposed under the PHMC model and the PHPH model. The proposed methods allow researchers to estimate the covariate effects separately and identify essential factors associated with the cure rate and the risk of failure. Simulation studies show that the proposed methods perform reasonably under finite sample sizes. We also demonstrate the proposed methodology on two real applications involving survival data with a cure fraction.

In this work, for the MC model, we propose two different incidence regression estimators $\hat{\alpha}_{KM}^{PO}$ and $\hat{\alpha}_{NP}^{PO}$. For the BCH model, we consider two different regression estimators $\hat{\gamma}_{KM}^{PO}$ and $\hat{\gamma}_{NP}^{PO}$ for *long-term* effects and two different regression estimators $\hat{\phi}_{KM}^{PO}$ and $\hat{\phi}_{NP}^{PO}$ for *short-term* effects. In the simulation studies, each of the pairs of estimators performs similarly. However, from the computing time point of view, we recommend that researchers consider estimators based on KM estimator as its computing times is fast. On the other hand, as we showed in the simulation studies, our proposed estimators based on pseudo-observations under both MC and BCH models have larger ESE and SEE than those from the usual MC and BCH models. This efficiency loss is commonly seen in the pseudo-observations literature.^{12,40} As pointed out by a referee, the efficiency loss is a limitation for our proposed pseudo-observations approach on the cure models when the number of covariates is small, like our real data analysis in Section 6.1. However, modeling pseudo-observations constitutes a general and straightforward approach to simplify survival analysis. In our applications, the pseudo-observations approach brings several advantages. First, it is more flexible and feasible for parameter estimation and variable selection when the number of covariates is large. The estimating procedure can be performed separately for each component in the MC model and each effect in the BCH model. Second, the computing time based on the pseudo-observations approach is faster than standard approaches that use the EM algorithm for estimation and bootstrapping for standard errors. Finally, pseudo-residuals can be applied for model diagnosis via residual plots.

Our PHPH model links the *short-term* effect to the baseline survival function via a proportional hazard model. The proposed PHPH model can be extended to accommodate non-proportional hazard assumptions by considering an exponential transformation or an accelerated failure time model in the *short-term* effect.⁸ On the other hand, we applied the pseudo-observations approach to the PHMC model. Another commonly used MC model is the accelerated failure time MC (AFTMC) model,⁷ in which the logistic regression model is considered to model the cure status in the incidence component and the AFT model, $\log(T) = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon$, is used to model the conditional survival function in the latency component, where ε is the error term with survival distribution $S_\varepsilon(\cdot)$. Under AFTMC model, our proposed pseudo-observations (6) and (7) using GEE approach can be directly applied to estimate the unknown parameters in the incidence component. For the latency component, we discuss the feasibility of using the pseudo-observations approach to estimate $\boldsymbol{\beta}$ under AFT model in two folds. First, when the survival distribution $S_\varepsilon(\cdot)$ is known; that is, a parametric AFT model, one can write $S_\varepsilon^{-1}(S_u(t)) = \log(t) - \mathbf{X}^T\boldsymbol{\beta}$ and treat it as a GLM with link function $S_\varepsilon^{-1}(\cdot)$. To estimate $\boldsymbol{\beta}$, one can create the pseudo-observations for $S_u(t)$ with given time points $t \in \{t_1, \dots, t_H\}$ based on KM estimator as we proposed in equation (8) and then the GEE approach can be applied to obtain the estimates for $\boldsymbol{\beta}$. However, one might need to program the estimating equation on their own because the geese function in the R package *geepack*³⁵ only provides commonly seen link functions. Second, when the survival distribution $S_\varepsilon(\cdot)$ is unknown in a semi-parametric AFT model, it may not be straightforward to use the pseudo-observations approach to estimate $\boldsymbol{\beta}$ even though the pseudo-observations for $S_u(t)$ can be created. The main issue is the unknown survival function $S_\varepsilon(\cdot)$, so is the inverse function $S_\varepsilon^{-1}(\cdot)$. Further investigation is required and will be an interesting topic for future research.

We note some limitations of our proposed work for variable selection. Based on our simulation studies, the TP rate seems to be low for the incidence component of the MC model when $n = 200$. This results might be induced because only one time point is used to create the pseudo-observations for the cure rate. When applying our proposed approach to a small sample sizes like 200 in our simulation, researchers can investigate the stability of variable selection via the bootstrap approach proposed by Royston and Sauerbrei.⁴¹ In this work, we only report the selected coefficient estimates from the penalized approach for the dental data analysis. In practice, one might adopt a two-stage approach in which the inference is based on the selected model to obtain standard errors.²⁷ The validity of inferences based on the penalization and the goodness-of-fit assessments will be studied in future work.

Acknowledgements

The authors would like to thank the editors, the associate editor, and three anonymous referees whose comments led to a substantial improvement of this paper. For this research work, Robert Platt acknowledges the support by the Natural Sciences and Engineering Research Council of Canada (NSERC) with grant numbers RGPIN-2016-06296 and RGPIN-2017-04363, the Canadian Institutes for Health Research (CIHR) with grant number FDN-143297 and the Albert Boehringer I Chair in Pharmacoepidemiology. Feng-Chang Lin acknowledges the support by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, USA, through grant number UL1TR002489.


Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.


Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR002489.

ORCID iDs

Chien-Lin Su  <https://orcid.org/0000-0002-0664-8583>

Sy Han Chiou  <https://orcid.org/0000-0003-2672-662X>

Feng-Chang Lin  <https://orcid.org/0000-0002-2638-1775>

Supplemental materials

The reader is referred to the online Supplemental Materials for asymptotic properties and additional simulation results. The R codes with examples for this work are deposited to a Github repository <https://github.com/stc04003/pseudo-cure>.

References

1. Kirkwood J, Strawderman M, Ernstoff M, et al. Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the eastern cooperative oncology group trial est 1684. *J Clin Oncol* 1996; **14**: 7–17.
2. Calhoun P, Su X, Nunn M, et al. Constructing multivariate survival trees: the MST package for R. *J Stat Softw* 2018; **83**: 1–21.
3. Peng Y and Yu B. *Cure Models: Methods, Applications, and Implementation*. New York: CRC Press, 2021.
4. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982; **38**: 1041–1046.
5. Peng Y, Dear K and Denham J. A generalized F mixture model for cure rate estimation. *Stat Med* 1998; **17**: 813–830.
6. Peng Y and Dear K. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000; **56**: 237–243.
7. Li CS and Taylor JMG. A semi-parametric accelerated failure time cure model. *Stat Med* 2002; **21**: 3235–3247.
8. Tsodikov A. Semi-parametric models of long-and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage. *Stat Med* 2002; **21**: 895–920.
9. A, Asselain B, Bardou V, et al. A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et Analyse de Donnees SpatioTemporelles* 1993; **12**: 66–82.
10. Sposto R. Cure model analysis in cancer: an application to data from the children's cancer group. *Stat Med* 2002; **21**: 293–312.
11. Tsodikov A, Ibrahim JG and Yakovlev AY. Estimating cure rates from survival data: an alternative to two-component mixture models. *J Am Stat Assoc* 2003; **98**: 1063–1078.
12. Andersen P, Klein J and Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multistate models. *Biometrika* 2003; **90**: 15–27.
13. Klein J, Logan B, Harhoff M, et al. Analyzing survival curves at a fixed point in time. *Stat Med* 2007; **26**: 4505–4519.
14. Andrei AC and Murray S. Regression models for the mean of the quality-of-life-adjusted restricted survival time using pseudoobservations. *Biometrics* 2007; **63**: 398–404.
15. Nicolaie MA, Van Houwelingen JC, DeWitte TM, et al. Dynamic pseudo-observations: a robust approach to dynamic prediction in competing risks. *Biometrics* 2013; **69**: 1043–1052.

16. Pavlič K and Perme MP. Using pseudo-observations for estimation in relative survival. *Biostatistics* 2019; **20**: 384–399.
17. Sabathé C, Andersen P, Helmer C, et al. Regression analysis in an illness-death model with interval-censored data: a pseudo-value approach. *Stat Methods Med Res* 2020; **29**: 752–764.
18. Su CL, Platt R and Plante JF. Causal inference for recurrent event data using pseudo-observations. *Biostatistics* 2022; **23**: 189–206.
19. Graw F, Gerds T and Schumacher M. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Anal* 2009; **15**: 241–255.
20. Jacobsen M and Martinussen T. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian J Stat* 2016; **43**: 845–862.
21. Overgaard M, Parner E and Pedersen J. Asymptotic theory of generalized estimating equations based on jack-knife pseudoobservations. *Ann Stat* 2017; **45**: 1988–2015.
22. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B (Methodological)* 1996; **58**: 267–288.
23. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006; **101**: 1418–1429.
24. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
25. Liu X, Peng Y, Tu D, et al. Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Stat Med* 2012; **31**: 2882–2891.
26. Masud AA, Yu Z and Tu W. Variable selection and nonlinear effect discovery in partially linear mixture cure rate models. *Biostat Epidemiol* 2019; **3**: 156–177.
27. Masud A, Tu W and Yu Z. Variable selection for mixture and promotion time cure rate models. *Stat Methods Med Res* 2018; **27**: 2185–2199.
28. Wang L, Zhou J and Qu A. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* 2012; **68**: 353–360.
29. Tsodikov A. A proportional hazards model taking account of long-term survivors. *Biometrics* 1998; **54**: 1508–1516.
30. Maller RA and Zhou S. Estimating the proportion of immunes in a censored sample. *Biometrika* 1992; **79**: 731–739.
31. Kaplan E and Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958; **53**: 457–481.
32. Tsodikov A. Estimation of survival based on proportional hazards when cure is a possibility. *Math Comput Model* 2001; **33**: 1227–1236.
33. Andersen P and Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res* 2010; **19**: 71–99.
34. Liang KY and Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.
35. Halekoh U, Højsgaard S and Yan J. The R package geeppack for generalized estimating equations. *J Stat Softw* 2006; **15**: 1–11.
36. Klein J, Gerster M, Andersen P, et al. SAS and R functions to compute pseudo-values for censored data regression. *Comput Methods Programs Biomed* 2008; **89**: 289–300.
37. Klein J and Andersen P. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* 2005; **61**: 223–229.
38. Cai C, Zou Y, Peng Y, et al. smcure: an R-package for estimating semiparametric mixture cure models. *Comput Methods Programs Biomed* 2012; **108**: 1255–1260.
39. Wang W. intsurv: Integrative Survival Models, 2019. URL <https://github.com/wenjie2wang/intsurv>. R package version 0.2.1.
40. Andersen P, Syriopoulou E and Parner ET. Causal inference in survival analysis using pseudo-observations. *Stat Med* 2017; **36**: 2669–2681.
41. Royston P and Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Stat Med* 2003; **22**: 639–659.