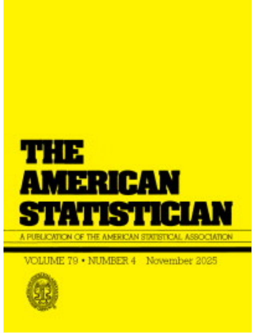

THE AMERICAN STATISTICIAN
A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION
VOLUME 79 • NUMBER 4 November 2025

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: www.tandfonline.com/journals/utas20

Feature selection in Cox model with partially observed covariates: Application to oncology trials

Ujjwal Das & Ranojoy Basu

To cite this article: Ujjwal Das & Ranojoy Basu (22 Dec 2025): Feature selection in Cox model with partially observed covariates: Application to oncology trials, The American Statistician, DOI: 10.1080/00031305.2025.2606077

To link to this article: https://doi.org/10.1080/00031305.2025.2606077

View supplementary material

Accepted author version posted online: 22 Dec 2025.

Submit your article to this journal

Article views: 43

View related articles

View Crossmark data

Feature selection in Cox model with partially observed covariates: Application to oncology trials

Ujjwal Das[1,*] and Ranojoy Basu[2]

[1]OM, QM & IS Area, IIM Udaipur, Rajasthan, India

[2]Economics and Development Studies Area, IIM Udaipur, Rajasthan, India

[*]*email:* ujjwal.das@iimu.ac.in

Abstract

In many real-life experiments with human subjects, missing data are common. Multiple imputation is widely used to handle unobserved data points. In statistical research, selecting important variables from multiple imputed datasets can be challenging, as each imputed data set may yield different sets of variables. Over the last decade, stacking imputed datasets and analyzing the resulting integrated data has gained attention. In this article, we consider both horizontal and vertical stacking approaches. The horizontal stacking approach in conjunction with different group penalties is discussed alongside the recently proposed vertical appending method, for identifying predominant variables under time-to-event data. The proposed methods are investigated numerically. Finally, the methods are illustrated in two real-world oncology experiments.

Keywords: Time to event data; missing covariate; missing at random; multiple imputation; stacking imputed datasets

1 Introduction

The Cox Proportional Hazards Model is a well-established method for analyzing survival time data. In recent decades, we have found its application in various domains, including health sciences, economics, finance, and operations research (Tsikriktsis (2005), Lee et al., (2021)). Complete covariate data for each subject are often unavailable in real-world scenarios such as clinical trials, observational studies, and financial stock market data. Missing data may arise due to many circumstances. In clinical trials, they may result from loss of hospital records, unavailability of covariate measurements, absence of biomarker information from participants, or survey non-response (Herring and Ibrahim (2001), Carroll et al., (2020)). A recent review on observational time-to-event studies in oncology trials reports the frequent occurrence of missing data in explanatory variables (using Medline and Embase databases, spanning from $2012-2018$, see Carroll et al., (2020) for further details). Missing values may arise for financial stock market data due to markets being closed for holidays, inability to capture financial data in the specified period, or recording errors (Lo and MacKinlay (1990), Bondon and Bahamonde (2012)). Missing observations also arise when jointly modeling financial and macroeconomic variables measured at different frequencies (Creal et al., (2014), Blasques et al., (2021)). Missing data may result in biased estimates and inaccurate statistical inference.

To handle missing data, one must assume the underlying missing mechanism. A commonly made assumption about the missing mechanism by practitioners is missing at random or MAR, where the missingness probability depends on the observed information (see, e.g., Li and Stuart (2019), Carroll et al., (2020)). Several approaches are available in the literature for dealing with incomplete covariates in the Cox model under the MAR assumption. Martinussen (1999), Herring and Ibrahim (2001) considered an EM-algorithm-based approach to handle missing covariates under MAR. Later, Garcia et al., (2010) introduced adaptive Lasso (ALASSO) and SCAD to perform feature selection on the EM adjusted log-likelihood. However, this method has not been implemented in any popular statistical software. Almost concurrently, Van Buuren et al., (1999) adopted multiple imputation (MI) under a similar framework. Since then, due to its ease of understanding and implementation in statistical software (like R, STATA, SAS), MI has evolved as the widely used method for treating missing data across various fields.

MI has been documented as the best general method for handling missing data. Table 3 of Carroll et al., (2020) shows that around 22% of clinical trials in oncology have adopted multiple imputation as one of the primary methods to handle missing data. In a recent paper, Enders (2023) shows the necessity and importance of handling and analyzing missing data with MI in the field of psychometry. MI is a two-stage process. The first stage involves replicating the original data multiple times. Next, in each replication, we replace missing values with plausible observations drawn from the appropriate distribution (Little and Rubin (2019)). The benefit of MI is that in addition to restoring the natural variability of the missing values, it incorporates uncertainty due to the missing data, which results in a valid statistical inference.

Several methods are available in the literature to address imputation. In this article, we focus on two different imputation methods. We first consider the well-known predictive mean matching (PMM), which is an attractive method to perform MI for missing data (Little and Rubin (2019), and White et al., (2011)). Following the influential paper of Van Buuren et al., (1999) on the practical use of MI, in a seminal paper, White and Royston (2009) developed a more principled approach to include a survival outcome in an imputation model. Furthermore, Van Buuren (2018) advocated the use of PMM approach of multiple imputation in a Cox model. More recently, Hsu and Yu (2018) has introduced a nonparametric MI method in a Cox model. We consider both MI approaches to include the partially observed covariates. Furthermore, the current practice is to analyze each imputed data and combine them using Rubin's Rule (Little and Rubin (2019)). However, if one tries to perform the Wald-type inference on imputed data, it may result in MLE's non-existence due to separation. Chen et al., (2009) developed a Monte Carlo EM (MCEM) algorithm to handle separation under MAR. However, the implementation of their proposal in software is not straightforward. Currently, a data with separation is discarded from further analysis following MI, and is simply replaced by another imputed dataset. Moreover, Meng (1994) observed that separate analysis of imputed data sets creates an incompatibility between the imputation model and the subsequent analysis procedures.

We propose two methodologies to address the challenges of analyzing imputed data in a Cox model. The first methodology involves stacking the imputed datasets horizontally and then applying group penalties to identify active factors. Our second methodology involves vertical stacking of the imputed data sets following the pioneering works of Chan (2022) and Chan and Meng (2022).

The article is structured as follows. Section 2 provides a brief overview of the Cox model, MI, group penalties, and the process of determining the tuning parameters after horizontal stacking. It also covers the methodology introduced by Chan (2022) and Chan and Meng (2022) for vertical stacking. Section 3 presents a numerical investigation of the performance of the proposed methods. In Section 4, we demonstrate our methods using an oncology trial dataset. Finally, Section 5 concludes the article. All proposed methods are implemented using R software (R core team (2022)), and the implementation is available as supplementary material.

2 Methods

Consider time-to-event data from *n* independent subjects. Let T^f denote the failure time, and *C* the censoring time with the additional assumption that T^f and *C* are independent, given the $p-$dimensional covariate information $\boldsymbol{X}$. Then $T = min(T^f, C)$ represents the observed time, with $\delta = I[T^f < C]$ denoting the censoring indicator. Hence, the observed data for the i^{th} individual is usually represented by the triplet $(T_i, \delta_i, \boldsymbol{X}_i)$, $i = 1, 2, \ldots, n$. The Cox Proportional Hazards Model with coefficient vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ can be written as

$$\lambda(t \mid \boldsymbol{X}) = \lambda_0(t) exp(\boldsymbol{\beta}' \boldsymbol{X}) \quad (1)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function.

Suppose that for any subject *i*, the corresponding covariate $\boldsymbol{X}_i$ is partially observed. In other words, for some subjects, there are instances of missing data. Further, we assume that missingness depends on survival outcome (i.e., survival time and censoring indicator) along with some fully observed covariates. This missing mechanism is known as missing at random (MAR, Little and Rubin (2019)). Several methods such as the EM-algorithm (Martinussen

(1999) Herring and Ibrahim (2001)) and multiple imputation (MI, Van Buuren et al., (1999), White and Royston (2009)) have been proposed in the literature to handle missing covariates in the Cox model. We consider the widely used MI, where all unobserved data are imputed multiple times (say, *M*) using a conditional model of the observed data. This method leads to the creation of *M* complete datasets, where fully observed variables remain the same, but partially observed predictors have different imputed values in missing places. The next section discusses the MI procedure.

2.1 Multiple imputation

Over the last few decades, MI has been a method of choice for handling missing data for two reasons: (i) its ease of implementation in several software and (ii) separate handling of missingness and its analysis. Although several algorithms are available, we adopt the *PMM* algorithm from MI by chain equations, as implemented in the R-package *mice* (Van Buuren & Groothuis-Oudshoorn (2011)). *PMM* can be used to impute continuous and discrete variables. It only assumes that the distribution of unobserved data is the same as the incumbents' observed data, making *PMM* more robust for imputation than other model-based methods. For a detailed discussion of the *PMM* algorithm and its limitations, we refer to Section 3.4 of Van Buuren (2018). To impute missing values in the covariates under the Cox model along with the observed data, we also include the Nelson-Aalen estimate of baseline hazard and censoring indicator as suggested in the literature (White and Royston (2009), Section 9.1.8 of Van Buuren (2018)). Thus, we generate M complete datasets where the columns without missing remain the same, but the partially observed columns contain different imputed values.

More recently, Hsu and Yu (2018) introduced a bootstrap-based nonparametric MI method, which is claimed to be less sensitive to misspecification of covariates in imputation compared to the PMM method. A linear/ generalized linear model is fitted to the observed values of the covariate containing missing data with survival outcomes and fully observed predictors as independent variables. This generates a predictive score that summarizes the relation between the missing covariate and the predictors of the model. Then a logistic regression is fitted with the missing mechanism with survival outcomes and fully observed predictors as independent variables, and a similar score is obtained. These models are fitted on a nonparametric bootstrap sample of the original dataset. For every subject that contains missing values, two scores are generated and used to select an imputing set consisting of similar subjects. These nearest neighbors (similar subjects) are used to impute the unobserved data by randomly drawing an observation. So, the procedure for imputation utilizes subjects fully observed for that variable. The process is repeated *M*-times to generate *M* copies of the data. The procedure is implemented in the R-package *NNMIS*. After imputation, we follow the same steps (horizontal/ vertical stacking) to obtain one dataset as discussed above.

For variable selection, the standard approach (implemented in several software like R [Van Buuren & Groothuis-Oudshoorn (2011)], SAS [SAS User's Guide (2004)], and Stata [White and Royston (2009)]) analyzes each *copy* of the datasets using the same complete-data method. This results in *M* sets of estimates and standard errors, which are combined following Rubin's rule (Little and Rubin (2019)). We adopt automatic variable selection techniques such as LASSO, SCAD and MCP for feature selection after MI. Previously, Garcia et al., (2010) adopted SCAD and adaptive LASSO for variable selection under EM-algorithm, but it does not create multiple datasets like MI. The application of these automatic approaches on *M* imputed datasets will likely result in the selection of different predictors from different datasets. When predictors originate from multiple datasets, deriving a single unified decision rule becomes challenging. For a complete dataset, the penalized objective function is expressed as

$$\min_{\boldsymbol{\beta}}\{L(\boldsymbol{\beta}|\ \mathbf{X}) + \sum_{j=1}^{p} P_{\lambda}(\beta_j)\},\ (2)$$

where $L(\boldsymbol{\beta}|\ \mathbf{X})$ denotes the Cox partial likelihood, λ is a tuning parameter, and $\sum_{j=1}^{p} P_{\lambda}(\beta_j)$ represents the penalty function (e.g., LASSO).

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M$ denote *M* imputed datasets, and let

$$\boldsymbol{\beta}_j = (\hat{\beta}_{1j}, \hat{\beta}_{2j}, \ldots, \hat{\beta}_{pj})$$

be the estimated regression coefficients from the *j*-th imputed dataset. Consider the *l*-th column of $\mathbf{X}$, denoted by $\mathbf{X}_l$. Ideally, if $\mathbf{X}_l$ is an active feature, we expect $\hat{\beta}_{lj} \neq 0$ for all $j = 1, 2, \ldots, M$; conversely, if $\mathbf{X}_l$ is inactive, we expect $\hat{\beta}_{lj} = 0$ for all *j*.

In practice, this consistency across imputations may not hold when penalties are applied separately to each imputed dataset. To address this, our first proposed method horizontally stacks the imputed datasets and applies the group penalties described in Subsection 2.2 to the combined data matrix

$$\mathbf{X}_{[1:M]} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M),$$

where the columns $(\mathbf{X}_{l1}, \mathbf{X}_{l2}, \ldots, \mathbf{X}_{lM})$ for $l = 1, 2, \ldots, p$ are treated as groups. For the second proposed approach, we vertically append the resulting datasets. We apply the likelihood ratio tests (LRTs) under the Cox model using the vertically appended data covered in Subsection 2.3.

2.2 Variable selection after horizontal stacking

Group LASSO: Before moving to grouped LASSO, we will briefly discuss the simple LASSO penalized regression problem. LASSO is introduced in Tibshirani (1996), which selects a sparse variable set by solving equation 2 with penalty $\boldsymbol{P}_{\lambda}(\beta) = \lambda|\beta|$. Using the singularity property of the L_1 norm at 0, LASSO yields exactly zero estimates for some of the estimated beta coefficients. Yuan and Lin (2006) proposed the grouped version of the LASSO problem with penalty $\boldsymbol{P}_{\lambda}(\|\boldsymbol{\beta}_j\|_2; a, \lambda) = (\sum_{k=1}^{M} \boldsymbol{\beta}_k^2)^{\frac{1}{2}}$. From the *p* predictors, we have *M* different groups. Chen and Wang (2013) proposed using group-LASSO for variable selection in linear regression with combined MI datasets. However, this method, known as MI-LASSO, has a limitation due to its use of the local quadratic approximation to the L_1-penalty. This approximation effectively turns the objective function into a ridge regression problem, which requires a manual threshold to set small coefficient estimates to zero. In an unpublished manuscript, Goeman et al., (2012) examined MI-LASSO for the Cox model with partially observed covariates, using a quasi-Newton method to maximize the objective function. More recently, Du et al., (2022) overcame this limitation by developing cyclic coordinate descent algorithms for grouped LASSO, elastic net, and their adaptive versions for binary and continuous outcomes. We extend this approach by incorporating other group-based penalties, such as SCAD and MCP, in addition to LASSO, to select important variables. The implementation of all these penalties uses the coordinate descent algorithm available in the R-package *grpreg*.

Group-SCAD: Like group-LASSO, we first introduce the non-group version of the SCAD penalty. For strictly non-convex penalty, (Fan and Li, 2001) introduced the SCAD penalty with the following penalty function

$$\boldsymbol{P}_{\lambda}(\beta;a,\lambda)=\begin{cases}\lambda|\beta|; & 0\le|\beta|<\lambda\\ \dfrac{\lambda a|\beta|-0.5(\beta^2+\lambda^2)}{a-1}; & \lambda\le|\beta|<a\lambda\\ (a+1)\dfrac{\lambda^2}{2}, & o.w.\end{cases} \quad (3)$$

Thus, the first derivative of the penalty term can be obtained as

$$\boldsymbol{P}_{\lambda}^{'}(\beta;a,\lambda)=\begin{cases}\lambda sgn(\beta); & 0\le|\beta|<\lambda\\ \lambda a sgn(\beta)-\beta; & \lambda\le|\beta|<a\lambda\\ 0, & o.w.\end{cases}$$

Fan and Li (2001) proposed parameter *a*'s value as 3.7, which we considered in this work. For group-SCAD, we minimize the objective function with the same penalty but β replaced by $\|\boldsymbol{\beta}\|_2=(\sum_{k=1}^{M}\boldsymbol{\beta}_k^2)^{\frac{1}{2}}$ group-norm for $k=1,\ldots,M$ groups, since, similar to group-LASSO, group-SCAD penalizes over group norms (Breheny and Huang (2015)).

Group-MCP MCP was first introduced in Zhang (2010). A general MCP solves a similar penalized optimization problem with penalty as

$$\boldsymbol{P}_{\lambda}(\beta;a,\lambda)=\begin{cases}\dfrac{2\lambda a|\beta|-\beta^2}{2a}, & \text{if } |\beta|\le a\lambda\\ \dfrac{\lambda^2 a}{2}, & \text{if } |\beta|>a\lambda\end{cases} \quad (4)$$

and the corresponding first derivative of penalty

$$\boldsymbol{P}'_{\lambda}(\beta; a, \lambda) = \begin{cases} \dfrac{\lambda a \operatorname{sign}(\beta) - \beta}{a} & \text{if } |\beta| \le a\lambda \\ 0 & \text{if } |\beta| > a\lambda \end{cases}$$

for some $a > 1$ and $\lambda \ge 0$. A common choice for the value of the parameter *a* is 3 (which is considered in this research). For group-MCP, we minimize the objective function with a similar group-norm $\|\boldsymbol{\beta}\|_2$ as group-LASSO and group-SCAD for $j = 1, \ldots, p$ groups and with penalty defined in 4.

Selection of Tuning Parameter: For all group penalties, selecting the tuning parameter λ is crucial, as it controls the sparsity of the model. An improperly adjusted model leads to biased estimates that suffer from prediction performance. Various methods, including cross-validation (CV), generalized CV and information criteria (IC) have been discussed in the literature. We consider the corrected Akaiki's information criterion (cAIC, Hurvich and Tsai (1989)) to select the optimal λ. The final parameter estimates are obtained by minimizing the cAIC, defined as

$$cAIC = -2\log L(\boldsymbol{\beta}) + 2\frac{df_{\lambda}}{n} + 2df_{\lambda}\frac{(df_{\lambda}+1)}{(n - df_{\lambda}+1)},$$

where df_{λ} is calculated with respect to each group penalty. The primary reason for choosing cAIC is its asymptotic consistency in model selection. Thus, the probability that the selected model converges with the true model increases to 1 as the sample size increases.

2.3 Analysis after vertical stacking

The salient feature of MI performing the imputation and analyses separately has led to some incompatibility between the imputation model and the subsequent analysis procedure (Meng (1994)). Later, Xie and Meng (2017) indicated that this uncongeniality may be overcome by adopting likelihood-based inference. In the *stacked* approach Du et al., (2022) added penalized loss functions from each of the imputed datasets and maximized the resulting penalized objective function. However, their method is not theoretically grounded. Furthermore, the non-existence of maximum likelihood estimate(s) also obstructs the adoption of Rubin's rule (Little and Rubin (2019)) through the Wald test in some scenarios under the Cox model. As an alternative, LRT after MI emerges as a solution to draw statistical inference. Several approaches have been attempted to combine LRTs (see references in Chan (2022)). However, they face limitations, including the stringent assumption of equal odds of missing information (OMI), non-invariant to reparameterization, etc. Recently, Chan (2022) proposed a less restrictive approach called stacked-MI, which involves vertical appending the imputed datasets and calculating the Likelihood Ratio Test (LRT) to address the limitations of traditional methods. This proposal applies to the Wald test, the LR test, and the score test. In addition to the development of the theory for tests after vertical stacking of imputed datasets, the error rates of these methods for linear and logistic regression models were examined in Chan (2022). This data augmentation alleviates the problem of separation. For time-to-event data, the Wald test statistic can be undefined for some parameters if separation occurs in the imputed dataset, making LRT preferable. Following Chan's algorithms (see Chan (2022)), we developed MI-LRT (Multiple Imputation

Likelihood Ratio Test) for the Cox model to provide a reliable statistical test for time-to-event data analysis.

Let $l_s(\boldsymbol{\beta})$ denote the log-likelihood of the vertically stacked data. Then, LRT-based test statistic denoted by $\hat{D}$ is calculated as

$$\hat{D} = \frac{\hat{d}_s}{k(1+(1+\frac{1}{M})\hat{\mu}_r)} \quad (5)$$

where $\hat{d}_s = -2(l_s(\boldsymbol{\beta}_{0s}) - l_s(\boldsymbol{\beta}_s))$ is the LRT statistic obtained by fitting the null and unrestricted models to the vertically stacked data with $\boldsymbol{\beta}_{0s}$ and $\boldsymbol{\beta}_s$ as the maximum likelihood estimates under the two models; $\hat{\mu}_r$ is an estimate of the average OMI, and *k* denotes the number of parameters of interest. Following Chan (2022), we estimate the OMI $r_1, r_2, \ldots, r_k$ using inverse transformations, implemented in the R-package *stackedMI*. The limiting null distribution of $\hat{D}$ is approximated using a Monte Carlo method. For each $j \in \{1, 2, \ldots, k\}$ draw G_j^l from χ_1^2 and H_j^l from $\chi_{M-1}^2/(M-1)$ independently for $l = 1, 2, \ldots, N$. Given $r_1, r_2, \ldots, r_k$ generate *N* random replicates from

$$\mathcal{D}^l = \frac{\frac{1}{k}\sum_{j=1}^{k}(1+(1+\frac{1}{M})\hat{r}_j)G_j^l}{1+\frac{1}{k}\sum_{j=1}^{k}(1+\frac{1}{M})\hat{r}_j H_j^l} \quad (6)$$

The following steps describe the algorithm for obtaining the $\hat{D}$ in (5) and its probability distribution after stacking the imputed datasets row by row. If $X^1, X^2, \ldots, X^M$ denotes the *M* appropriately imputed datasets, let $X^{\{1:M\}}$ denote the vertically stacked *M* datasets and $X^{\{-l\}}$ denote the same without X^l.

• Calculate the LRT $\hat{d}_s$ statistic from the stacked dataset

• For $l \in \{1, 2, \ldots, M\}$

– Compute LRT statistic $\hat{d}^{\{-l\}}$ from $X^{\{-l\}}$

– From X^l compute LRT statistic $\hat{d}^l$

– Calculate $\hat{T}_l = \hat{d}^{\{-l\}} + \hat{d}^l - \hat{d}_s$

• Calculate $\hat{t}_j = \sum_{l=1}^{M} \frac{\hat{T}_l^{\,j}}{M}$

- Define $R_\tau = \sum_{j=1}^{k} r_j^\tau$ for $\tau = 1, 2, \ldots, k$; $t_1 = R_1$ and $t_\tau = \sum_{j=1}^{\tau} \frac{(\tau-1)!}{(\tau-j)!} 2^{(j-1)} R_j t_{\tau-j}$
- Invert the above functions to get an estimate $\hat{r}_j$ of r_j, for $j = 1, 2, \ldots, k$
- Obtain $\hat{D}$ from (5)
- For $j = 1, 2, \ldots, k$ get the limiting distribution of the above test statistic from (6)
- The p-value is the proportion of $\mathcal{D}^l$ exceeding $\hat{D}$

Similarly, Chan and Meng (2022) introduced an alternative MI-LRT statistic, which shows that using a robust estimation of the OMI makes the test less sensitive to the assumption of an equal OMI. We incorporate this robust LRT in our approach. Estimation of the test statistic resulting from LRT_R along with OMI (r^{rob}) can be performed as follows:

- Calculate $\bar{\delta} = \frac{2}{M} \Sigma_{i=1}^{M} l(\boldsymbol{\beta}^i, X^i)$ where $l(\boldsymbol{\beta}^i, X^i)$ is the log-likelihood and $\boldsymbol{\beta}^i$ is the maximum likelihood estimator from i^{th} imputed dataset X^i with $i = 1, 2, \ldots, M$

- $\hat{\delta}_s = \frac{2}{M} l_s(\boldsymbol{\beta})$ where $l_s(\boldsymbol{\beta})$ is the log-likelihood and $\boldsymbol{\beta}$ is the maximum likelihood estimator from the stacked data

- Estimate *r* by $r^{rob} = \frac{(M+1)}{h(M-1)(\bar{\delta} - \hat{\delta}_s)}$, where *h* denotes the number of parameters in the model

- Calculate the test statistic $\hat{D}_s$ given by

$$\hat{D}_s = \frac{\hat{d}_s}{k(1 + r^{rob})} \quad (7)$$

and degrees of freedom ν_s

$$\nu_s = h(M-1)\left[1 + \frac{1}{r^{rob}}\right]^2 \quad (8)$$

where *k* is the number of parameters of interest

- Compute the p-values as $P(\hat{D}_s > F_{k,\nu_s}^{-1}(1-\alpha))$, where $F_{k,\nu_s}^{-1}(1-\alpha)$ denotes the $(1-\alpha)$-th quantile of the *F*-distribution with numerator degrees of freedom *k* and denominator degrees of freedom ν_s

We observe that both $\hat{D}$ and $\hat{D}_s$ mainly rely on the stacked dataset and the pooled estimate of the LRT statistic, rather than on the pooling point estimates of the parameters. This approach makes them less prone to the separation issue found in some of the imputed datasets. For a detailed discussion on the construction of the test statistic, the estimation of missing information odds, and the limiting distribution of $\hat{D}$ and $\hat{D}_s$ refer to Chan (2022) and Chan and Meng (2022).

3 Numerical Investigation

This section examines numerically the performance of the proposed methods. The discussion on the type-I error rate of the proposed MI-LRTs is provided in the supplementary material for brevity. Here, we study the comparative performance of all the methods discussed in this paper.

3.1 Data generation

To remain consistent with the real data scenario, we consider $p = 16$ predictor variables with three different sample sizes $n = 50,100$ and 200. The design matrix *X* with *p* columns is generated from a multivariate normal distribution with mean vector 0 and unit variance. The correlation between the i^{th} and j^{th} columns is set to $\rho^{|i-j|}$ with $\rho = 0.5$. We consider 5 and 35 as missing percentages in some of the columns of *X*. The missing data in the covariates is generated using the *ampute* function from the *mice* package, with MAR as the missingness mechanism. Furthermore, for every combination of sample size and missing percentage, the proposed methods are evaluated in three different designs: (*i*) $\beta_j = 1$ for $j = 1, \ldots, 5$ and the rest are set to zero; (ii) $\beta_j = 1$ for $j = 9, \ldots, 13$ and the rest are set to zero; (iii) $\beta_j = 1$ for $j = 6, \ldots, 11$ and the rest are set to zero. Without loss of generality, missing observations are created in the last eight columns of *X* using the R-package *mice*. Collectively, we consider three different situations: Design 1, where all 5 active variables are completely observed; Design 2, where all 5 active variables are partially observed; and Design 3, where three of the six active variables are fully observed and the remaining three have missing information. Survival times t_i are generated from an exponential distribution with hazard $h_i = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{16} X_{16})$, where β_0 is chosen as 0.5. Survival times are randomly censored with a probability of 0.20. The entire process was replicated 1000 times in each scenario.

The imputations of the missing observations are performed following the suggestions in White et al., (2011). We used the Nelson-Allen estimate along with $X_1, \ldots, X_8$ with *PMM* option from *mice* (Van Buuren & Groothuis-Oudshoorn (2011)) for imputations. The number of imputations (M) is chosen as 100. The imputed columns are aligned side by side with the fully observed columns $X_1, \ldots, X_8$. This yields a "complete" design matrix with *n* rows and $(8 + 8 \times M)$ columns. We call the R-package *grpreg* with three penalties: group LASSO, group SCAD and group MCP. The columns $X_1, \ldots, X_8$ have group ID $1, \ldots, 8$, whereas $X_9, \ldots, X_{16}$ have M groups from M imputed copies. As discussed in Section 2, the tuning parameter λ is selected by minimizing cAIC.

For vertical stacking, we append the imputed datasets in the “long” format of the R-package *mice*. This yields data with *Mn* rows and 16 columns. Henceforth, the methods based on the likelihood ratio tests of Chan (2022) and Chan and Meng (2022) will be denoted by LRT_C and LRT_R, respectively. Considering the rows as independent, we call the R-function *coxph* to determine the $\hat{D}$ statistic, mentioned in Section 2. The imputation and the analyses are implemented in an R-function. We used several functions from the R-package *stackedMI* to estimate the OMIs and related expressions along with the p-values. Additionally, we consider MI of the data using the method proposed in Hsu and Yu (2018) using the R-package *NNMIS* and then stack the datasets in the same manner.

3.2 Comparison of methods after stacking

The methods are compared using three criteria: true model identification rate, abbreviated as *TMIR*, which is the proportion of time the entire correct model is detected; average number of correct variables identification rate, abbreviated as $AVG_C IR$, which is the correctly identified non-zero variables in the simulation; and the average number of false positives, abbreviated as *AVFP*, which is the average number of nonactive variables incorrectly detected as non-zero during the simulation. After vertical stacking, the important variables are identified at the 5% significance level. The results of the numerical studies are summarized in Tables 1, 2 and 3 for the three different designs. The best number within every metric is highlighted in the tables unless multiple methods achieve equal performance.

Table 1 shows that for sample sizes of 100 or larger, all group penalties after horizontally stacking multiple imputed datasets perform equally well. However, Tables 2 and 3 reveal that group LASSO identifies the full model more frequently and achieves a higher rate of correct variable detection, although it also includes slightly more false positives. The *TMIR* from *NNMIS* is often higher than that obtained with *PMM*. For sample sizes less than 100, group LASSO outperforms the other methods. When the sample size is 200, *TMIR* and $AVG_C IR$ from LRT_R improve significantly, with fewer *AVFP* across all designs and missing data percentages. Conversely, LRT_C performs less effectively, even with a sample size of 200. Across all simulations, LRT_C has the smallest *TMIR*. The observations for LRT_C and LRT_R align with the findings of Type-I error presented in the supplementary material. The conservative nature of LRT_R for $n \leq 200$ results in a lower *TMIR* than that of group LASSO. Therefore, horizontal stacking combined with group penalties and non-parametric imputation can be recommended for sample sizes of 200 or less. For larger sample sizes, vertical stacking, after appropriate imputation, along with LRT_C, generally selects a subset of important variables, while LRT_R performs similarly to other group-penalized methods.

We note the divergence of at least one parameter estimate to infinity in one or more imputed data while calculating $\hat{D}$ and $\hat{D}_s$, mainly for $n \leq 100$ in the simulations. However, the likelihood remains finite, and its convergence is attained. The phenomenon, known as separation, has been addressed by various methods in the literature to tackle this issue (see, e.g. Heinze and Schemper (2001)). Since our method utilizes the likelihood function instead of the point estimates, separation does not affect the analysis. Penalized methods are known to be resistant to separation; therefore, all of the methods proposed in the article are insensitive to it.

4 Real data analysis

We apply the proposed methods on two real world oncology datasets where some covariates were partially observed.

4.1 Liver cancer data analysis

We illustrate our proposed methods for screening active factors with survival outcomes in patients with colorectal liver metastases (CLM). Upfront surgery is used routinely to increase the overall survival of patients. Currently, a minimally invasive preoperative strategy named portal vein embolization (PVE) is adopted before hepatic resection to increase the size of liver segments (future liver remnant or FLR) that will remain after surgery. However, PVE sometimes results in some adverse effects, such as a higher rate of tumor progression, which can ultimately impact patient survival. Recently, Collin et al., (2019) compared the effect of PVE on subjects' long-term survival and cancer recurrence. This study includes 71 subjects who underwent PVE and 57 subjects who did not. The sample size is the largest among earlier studies to compare survival between PVE and non-PVE groups. Owing to the complexity of the treatment process, tumor progression, etc., subject dropout is common in such oncology trials (see, e.g., Chow and Chok (2019)).

We explore the impact of possible variables listed in Tables 4 and 5 of Collin et al., (2019), on overall survival (OS time) of patients. The values in "tumoral volumes percentages" are not available for patients who underwent upfront surgery; therefore, this variable was removed. Additionally, we removed a subject with information missing from 11 of 16 variables. The variables "Bilateral-disease" and "Extra-hepatic-metastasis" were missing for only one subject, which was also excluded. This finally results in seven variables: Age, PVE, total-number-of-liver-metastases, gender, prevalence of cardiovascular disease, Bilateral-disease and Extra-hepatic-metastasis without missing values from 125 subjects. The remaining 8 variables contain unobserved values ranging from 3.8% to 34.7%. Missing values for some of the variables occur mainly in subjects who underwent *PVE*. Furthermore, subjects with unobserved operative time and cardiovascular disease may be considered short-lived, since for both variables, the median survival time for partially observed subjects is lower than than that of completely observed subjects. Finally, survival rate (Kaplan-Meier) for partially observed subjects and completely observed subjects is shown in Figure 1, which clearly shows lower survival rate for partially observed subjects. Hence, it may be safe to assume that the predictors are missing at random. Next, we impute the data following White and Royston (2009) with $M = 100$. After horizontally and vertically stacking the imputed datasets, we analyze the data as discussed in Section 2. Table 4 summarizes the non-zero estimates from group LASSO, SCAD and MCP after horizontal stacking and variables with p-value < 0.05 from LRT_C and LRT_R after vertical stacking, for overall-survival times. Variables selected by MI-LRT are marked with $\surd$. In many situations "PMM" and "NNMIS" imputations included the same sets of variables. For comparison, we also performed complete case analysis (CCA). This leaves us with 56 subjects. Moreover, the predicted survival rates for a subject who underwent PVE, and for a subject without PVE from group LASSO, group SCAD, and group MCP along with CCA are displayed in Figures 2, 3 and 4. To illustrate the impact of different group penalties on survival curves between *withPVE* and *withoutPVE*, we selected two subjects. One of them is a 65.5-year-old male who has *PVE* with other covariates "TOTAL number of liver Metastases" $= 3$, no cardiovascular disease, no diabetes, no "Bilateral Disease", "Number of NeoAdj chemo Cycles" $= 0$ and "Blood-Losses"

$= 500ml$. The other is a 62.7-year-old female who does not have PVE with "TOTAL number of liver Metastases" $= 2$, no cardiovascular disease, no diabetes, no "Bilateral-Disease", "Number-of-NeoAdj-chemo-Cycles" $= 3$ and "Blood-Losses" $= 500ml$. Estimated survival curves of the group penalties are compared with those of *CCA*.

In Table 4, we note that group SCAD and group MCP identify many variables, whereas group LASSO selects only three: "Age", "Total number of liver metastases", and "Prevalence of at least one cardiovascular disease". Both group SCAD and group MCP do not select "Age". Both group SCAD and group MCP select "PVE" with a negative coefficient, implying a higher predicted survival rate for patients with "PVE" than for those without "PVE" (see Figures 3 and 4). However, Group LASSO does not include "PVE". Collectively, these findings suggest that CLM patients with "PVE" have lower hazard rates than those without "PVE", or that "PVE" has no meaningful effect on patient survival. We also note that "PVE" is not a statistically significant variable in Collin et al., (2019) nor in our CCA, which supports the claim that, at the very least, "PVE" does not adversely affect overall survival (see, e.g., Chow and Chok (2019) and references therein) for patients with CLM.

Furthermore, " LRT_C " identifies "Total number of liver metastases" and "Operative time" as important predictors. In addition to these two variables, " LRT_R " also recognizes "Blood loss" as an important biomarker for overall survival time. From Table 4, we note that "Total number of liver metastases" is, perhaps, the most important predictor that influences overall survival, followed by "Blood loss", since three of the methods identify this variable. We also find that "PVE", "Prevalence of at least one cardiovascular disease", and "Bilateral disease" are each selected by at least two of the three group selection methods. These biomarkers are clinically meaningful and can be regarded among the most important predictors of overall survival for CLM patients.

The results across methods are consistent with our numerical investigations (Section 3), which indicate that for sample sizes around $n \approx 100$, the " LRT_R " approach is more likely to fail to select the appropriate predictors than group penalties, and that group penalties tend to produce fewer false positives. "Operative time" may thus be considered a potential false positive from " LRT_R ". Finally, CCA identifies "Operative time", "Extended right resection", and "Blood loss" as significant predictors.

Our analysis indicates that group SCAD and group MCP select "PVE" with a negative coefficient estimate, which corresponds to a higher predicted survival rate for patients with "PVE" than for those without `PVE` (Figures 3 and 4). In contrast, group LASSO does not select `PVE` as a relevant predictor. Taken together, these results suggest that "PVE" is associated with a lower hazard rate or has no meaningful impact on overall survival in CLM patients. This finding is consistent with the observation that "PVE" was not a significant covariate in Collin et al., (2019) and in our CCA, supporting the view that `PVE` does not adversely affect overall survival (see, e.g., Chow and Chok, 2019 and references therein).

Figure 3 shows that the predicted survival probability of a subject with "PVE" is higher than that of a subject without "PVE" at a given time point. A similar pattern is observed in Figure 4. The survival curves look similar for "PVE" and no-"PVE" from Figure 2, possibly because group LASSO has included only three covariates in the model. The outcome of LRT_R is in tandem with Collin et al., (2019) that "PVE" does not make any difference in the overall survival of patients with CLM. Overall, one may conclude that both group penalties and MI-LRTs yield parsimonious models when analyzing the overall survival time of patients suffering from CLM. Similar results are obtained after non-parametric MI using the R-package *NNMIS*.

5 Concluding remarks

In this article, we explore two methods to identify key features in Cox regression models when some covariates have missing data. Assuming the data are missing at random, we consider vertically and horizontally stacking imputed datasets rather than merging the results from multiple imputed datasets. We also incorporate more advanced and statistically robust techniques, such as MI-LRT, to identify important predictors from the combined data. From our simulation studies, LRT_R performs better, mostly for larger sample sizes. Group penalties after horizontal stacking are preferred over the MI-LRTs for smaller sample sizes. These recommendations are valid regardless of missing percentages in the MAR mechanism. In the statistical literature, LRT–based variable selection is sometimes viewed as less formal than approaches that optimize an explicit objective function. This may be because the LRT does not directly minimize a metric such as prediction error or a model selection criterion like AIC (Klauer (2001)). In contrast, penalization methods such as LASSO implicitly target prediction error minimization, which can make them appealing in certain contexts.

We note that both proposed procedures are invulnerable to separation, which may be present in the original data or in one or more imputed datasets. Hence, we do not need to discard the imputed datasets that suffer from separation. This leads us to draw exactly 100 imputed datasets for all the situations considered here. In addition, both proposals can be easily implemented in statistical software such as R. Hence, we recommend our proposals for stacking the imputed datasets and then analyzing the combined data over the existing approach for selecting variables in the Cox regression model. Furthermore, it is easy to construct confidence intervals based on profile likelihood using the vertical stacking approach.

This research opens several new avenues, some of which we are currently pursuing. In recent years, multiple imputation has been increasingly used to address missing data in various survival models. We are focusing on implementing MI-LRTs with partially observed time-varying covariates, as discussed by Keogh and Morris (2018). The performance of LRT_C and LRT_R in accelerated failure time models may be examined, particularly under model misspecification and different missing data mechanisms. Ahn (2018) explored MI in the context of interval-censored predictors in the Cox model, while Beesley et al., (2016) investigated the cure model with Cox regression, where missing covariates were imputed using MI. In all of these studies, Rubin’s rule was used for inference after MI. We intend to apply our proposed methods to these scenarios in future research. We also plan to investigate the performance of the proposed methods in more complex designs as discussed in Chen and Yi (2020) and Chen and Yi (2021). Another natural extension of this work is to evaluate the performance of the proposed methods in ultrahigh-dimensional settings ($p \gg n$). Penalized approaches such as LASSO and SCAD can face challenges when the number of predictors greatly exceeds the sample size (Chen (2021)). In such scenarios, the recently proposed $C-index$ –based method of Chen (2024) provides a unified framework for screening active features.

Acknowledgment: The authors are grateful for constructive feedback from the associate editor and two anonymous reviewers that sufficiently improve the quality of the article. Furthermore, the authors thank Professor Patrick Royston for kindly sharing the kidney cancer data.

Declaration: The authors report that there are no competing interests to declare.

References

AHN S., LIM J., PAIK M. C., SACCO R. L. & ELKIND M. S. (2018). Cox model with interval-censored covariate in cohort studies. Biometrical Journal, 60(4), 797-814.

BEESLEY L. J., BARTLETT J. W., WOLF G. T. & TAYLOR, J. M. (2016). Multiple imputation of missing covariates for the Cox proportional hazards cure model. *Statistics in Medicine*, **35(26)**, 4701-4717.

BLASQUES, F., GORGI, P., & KOOPMAN, S. J. (2021). Missing observations in observation-driven time series models. *Journal of Time Series*, **221(2),** 542-568.

BONDON, P., BAHAMONDE, N. (2012). Least squares estimation of ARCH models with missing observations. *Journal of Time Series,* **33 (6),** 880–891.

BREHENY P. & HUANG J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, **25**, 173-187.

CARROLL, O. U., MORRIS, T. P., & KEOGH, R. H. (2020). How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review. *BMC Medical Research Methodology*, **20(1),** 1-15.

CHAN, K. W. (2022). General and feasible tests with multiply-imputed datasets. *The Annals of Statistics*, **50(2)**, 930-948.

CHAN, K. W., & MENG, X.L. (2022). Multiple improvements of multiple imputation likelihood ratio tests. *Statistica Sinica*, **32**, 1489-1514.

Chen, L. P. (2021). Feature screening based on distance correlation for ultra-high dimensional censored data with covariates measurement error. *Computational Statistics*, **36**, 857-884.

Chen, L. P. (2024). Feature screening via concordance indices for left-truncated and right-censored survival data. *Journal of Statistical Planning and Inference*, **232**, 106153.

CHEN, M. H., IBRAHIM, J. G., & SHAO, Q. M. (2009). Maximum likelihood inference for the Cox regression model with applications to missing covariates. *Journal of Multivariate Analysis*, **100(9)**, 2018-2030.

Chen, L. P., & Yi, G. Y. (2020). Model selection and model averaging for analysis of truncated and censored data with measurement error. *Electron. J. Statist.*, **14(2)**, 4054-4109.

Chen, L. P., & Yi, G. Y. (2021). Analysis of noisy survival data with graphical proportional hazards measurement error models. *Biometrics*, **77(3)**, 956-969.

Chen Q. & Wang S. (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*, **32(21)**, 3646-3659.

Chow F.C. & Chok K.S. (2019). Colorectal liver metastases: An update on multidisciplinary approach. *World J Hepatology*, **11(2)**, 150-172.

Collin, Y., Paré, A., Belblidia, A., Létourneau, R., Plasse, M., Dagenais, M., … & Vandenbroucke-Menu, F. (2019). Portal vein embolization does not affect the long-term survival and risk of cancer recurrence among colorectal liver metastases patients: a prospective cohort study. *International Journal of Surgery*, **61**, 42-47.

Creal, D., Schwaab, B., Koopman, S.J., Lucas, A., (2014). Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics*, **96(5),** 898–915.

Du J, Boss J, Han P, Beesley LJ, Kleinsasser M, Goutman SA, Batterman S, Feldman EL & Mukherjee B. (2022). Variable selection with multiply-imputed datasets: choosing between stacked and grouped methods. *Journal of Computational and Graphical Statistics*, **31(4)**, 1063-1075.

Enders C.K. (2023). Missing data: An update on the state of the art. *Psychological Methods*. doi: https://doi.org/10.1037/met0000563

Fan J. & Li R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96(456)**, 1348-1360.

Garcia R. I., Ibrahim J. G. & Zhu H. (2010). Variable selection in the cox regression model with covariates missing at random. *Biometrics*, 66(1), 97-104.

Goeman, J. J., Putter, H., & Meulman, J. J. (2012). The group lasso in the proportional hazards model with an application to multiply imputed high-dimensional data. *unpublished Master Thesis*; available at *https://www.universiteitleiden.nl/binaries/content/assets/science/mi/scripties/luijkmaster.pdf*.

Grund, S., Lüdtke, O., & Robitzsch, A. (2023). Pooling methods for likelihood ratio tests in multiply imputed data sets. *Psychological Methods*, **28(5)**, 1207-1221.

Heinze, G., & Schemper, M. (2001). A solution to the problem of monotone likelihood in Cox regression. *Biometrics*, **57(1)**, 114-119.

Herring, A. H. & Ibrahim, J. G. (2001). Likelihood-based methods for missing covariates in the Cox proportional hazards model. *Journal of the American Statistical Association*, **96(453)**, 292-302.

HSU, C. H. & YU, M. (2018). Cox regression analysis with missing covariates via nonparametric multiple imputation. *Statistical Methods in Medical Research*, **28(6)**, 1676-1688.

HURVICH, C.M. AND TSAI, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.

KEOGH R. H., & MORRIS T. P. (2018). Multiple imputation in Cox regression when there are time-varying effects of covariates. *Statistics in Medicine*, **37(25)**, 3661-3678.

KLAUER, K.C. (2001). Theory of Model Testing and Selection, Editor(s): Neil J. Smelser, Paul B. Baltes, *International Encyclopedia of the Social & Behavioral Sciences,* Pergamon, 9927-9931, ISBN 9780080430768, https://doi.org/10.1016/B0-08-043076-7/00599-4.

LEE, K. J., TILLING, K. M., CORNISH, R. P., LITTLE, R. J., BELL, M. L., GOETGHEBEUR, E., … & CARPENTER, J. R. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Clinical Epidemiology,* **(134),**, 79-88.

LI P. & STUART E. A. (2019). Best (but oft-forgotten) practices: missing data methods in randomized controlled nutrition trials. *The American Journal of Clinical Nutrition*, **109(3)**, 504-508.

LITTLE, R. J. A., AND D. B. RUBIN (2019). *Statistical Analysis with Missing Data.*, John Wiley & Sons, New York

LO, A.W., MACKINLAY, A.C. (1990). An econometric analysis of nonsynchronous trading. *Journal Econometrics*, **45**, 1-2.

MARTINUSSEN T. (1999). Cox regression with incomplete covariate measurements using the EM-algorithm. *Scandinavian Journal of Statistics*, **26**, 479–491.

MENG, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical science*, 538-558.

R CORE TEAM (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

SAS Institute Inc. SAS/STAT 9.1 User’s Guide, Chapter 46. SAS Institute Inc.: Cary, NC, 2004.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58(1)**, 267-288.

TSIKRIKTSIS, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management,* **96(5),**, 53-62.

Van Buuren S, Boshuizen H.C. & Knook DL. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18, 681–694.

VAN BUUREN S., & GROOTHUIS-OUDSHOORN K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, **45**, 1-67.

VAN BUUREN, S. (2018). Flexible imputation of missing data. CRC press.

WHITE I. R., ROYSTON P., & WOOD, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine,* **30(4)**, 377-399.

WHITE I. R. & ROYSTON P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine*, **28(15)**, 1982-1998.

XIE X. & MENG X. L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: what happens when god's, imputer's and analyst's models are uncongenial?. *Statistica Sinica*, 1485-1545.

YUAN M. & LIN Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **68(1)**, 49-67.

ZHANG C.H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38(2)**, 894-942.

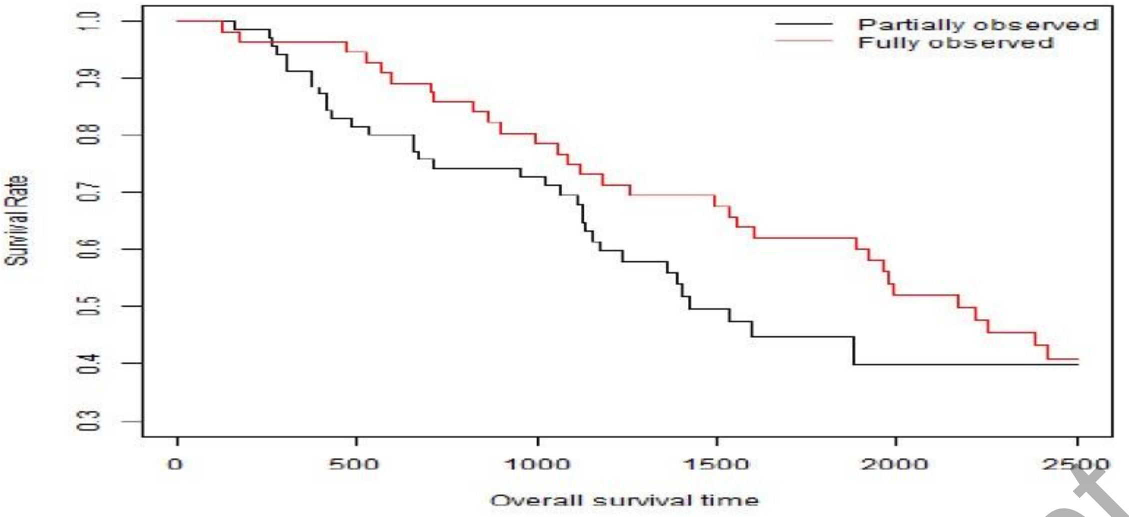

Figure 1: Survival rates from partially observed and fully observed subjects

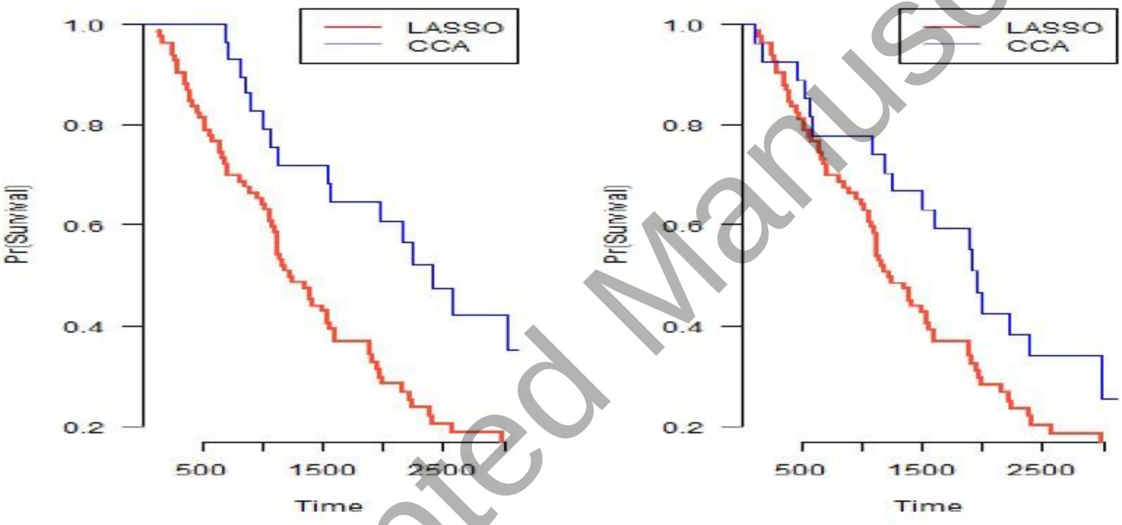

Figure 2: Predicted survival curves for a subject with PVE (left panel) and without PVE (right panel), obtained from the Group LASSO and complete-case analyses.

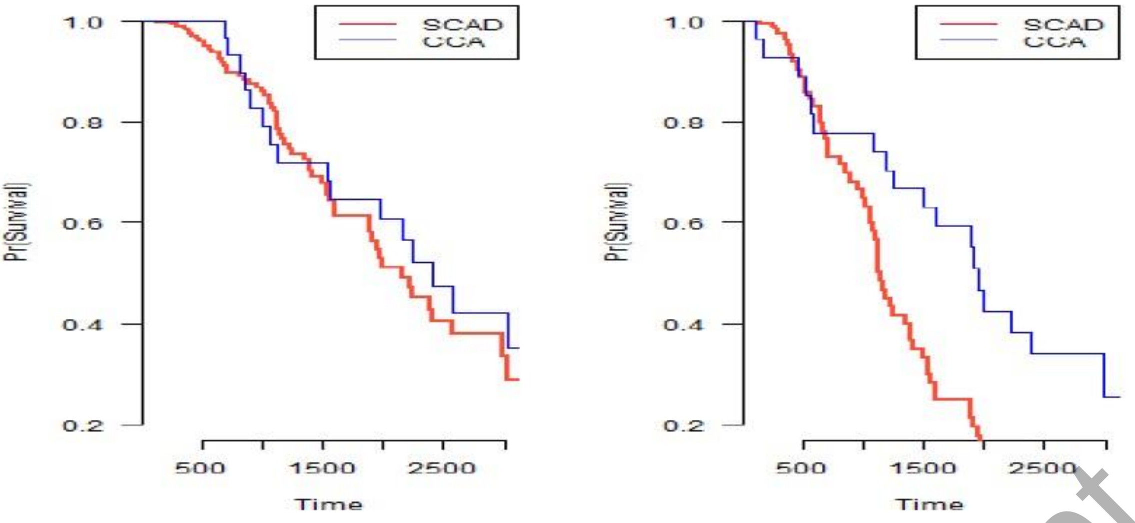

Figure 3: Predicted survival curves from the Group SCAD model (blue) and the complete-case analysis (red) for subjects with PVE (left panel) and without PVE (right panel).

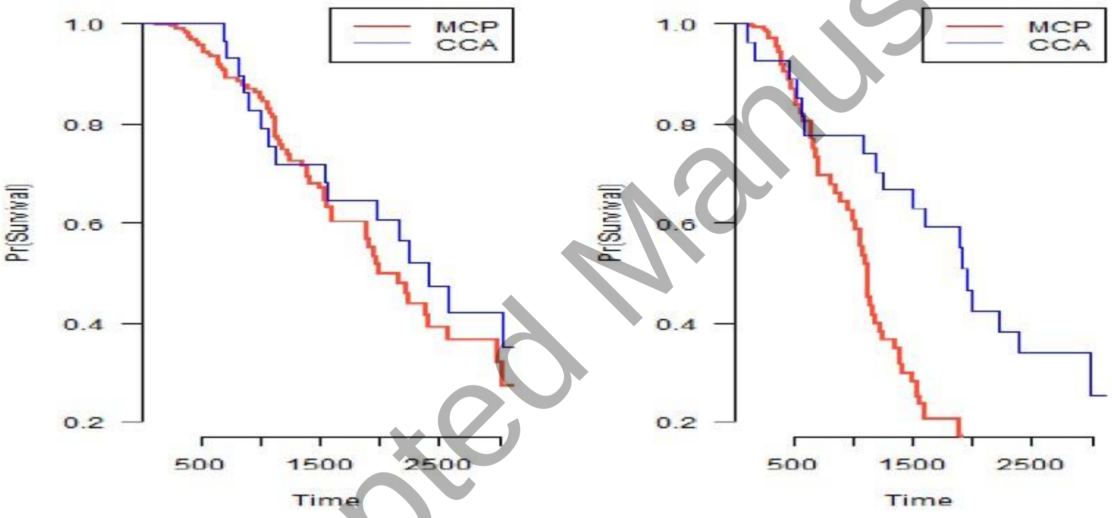

Figure 4: Predicted survival curves from the Group MCP model (blue) and the complete-case analysis (red) for subjects with PVE (left panel) and without PVE (right panel).

Table 1: Simulation Result for Design 1: All 5 active variables are fully observed and remaining 11 have missing values; *TMIR* closer to 100 is better; $AVG_C IR$ closer to 5 is better; *AVFP* closer to 0 is better

n	Miss%	Method	TMIR					$AVG_C IR$					AVFP				
			L_1	SCAD	MCP	LRT_C	LRT_R	L_1	SCAD	MCP	LRT_C	LRT_R	L_1	SCAD	MCP	LRT_C	LRT_R
50	5	PMM	55.1	44.6	44.1	14.8	33.2	**4.09**	3.87	3.65	1.94	2.57	3.19	4.36	3.39	0.12	2.44
		NNMIS	**56.2**	44.9	42.8	14.6	33.8	4.05	3.79	3.66	1.97	2.55	3.20	4.32	3.81	**0.11**	2.81
	35	PMM	53.4	32.2	32.6	10.8	25.1	3.98	3.63	3.64	1.86	2.08	6.05	4.94	3.44	0.14	2.70
		NNMIS	**54.8**	33.4	33.2	11.6	26.9	**4.13**	3.64	3.62	1.68	2.04	5.97	4.98	3.48	**0.12**	2.71
100	5	PMM	**79.2**	75.2	76.5	12.1	65.5	**4.51**	4.48	4.50	2.94	3.62	3.39	3.00	1.54	**0.92**	1.43
		NNMIS	79.1	75.8	76.2	11.9	65.4	4.50	4.48	4.49	2.95	3.69	2.58	2.33	1.25	1.04	1.01
	35	PMM	79.1	69.6	68.0	12.2	60.6	4.45	4.22	4.20	2.76	3.48	5	3.48	1.42	1.02	1.14
		NNMIS	**79.9**	69.4	67.8	12.0	60.5	**4.48**	4.24	4.22	2.74	3.46	4.3	3.01	1.98	**0.99**	1.12
200	5	PMM	100	100	100	15.6	100	5	5	5	3.25	5	2.8	2.4	1.5	0.95	1.04
		NNMIS	100	100	100	15.8	100	5	5	5	3.26	5	2.6	2.5	1.6	**0.94**	1.03
	35	PMM	100	100	100	14.6	92.0	5	5	5	3.08	4.85	2.6	2.4	1.5	0.95	1.05
		NNMIS	100	100	100	14.8	92.1	5	5	5	3.10	4.86	2.5	2.4	1.4	**0.94**	1.02

Table 2: Simulation Result for Design 2: All five active variables contain at least one missing values; *TMIR* closer to 100 is better; $AVG_C IR$ closer to 5 is better; *AVFP* closer to 0 is better

n	Miss %	Method	TMIR					$AVG_C IR$					AVFP				
			L_1	SCAD	MCP	LRT_C	LRT_R	L_1	SCAD	MCP	LRT_C	LRT_R	L_1	SCAD	MCP	LRT_C	LRT_R
50	5	PMM	49.1	39.2	32.5	14.2	34.2	**3.43**	3.16	2.04	1.92	2.54	6.25	4.69	2.71	0.16	2.96
		NNMIS	**49.8**	38.9	32.6	14.3	34.4	3.41	3.25	2.08	1.98	2.53	5.97	4.26	2.68	**0.15**	2.94
	35	PMM	37.7	34.2	32.2	11.2	25.4	3.18	3.22	2.05	1.89	2.57	6.33	5.64	3.124	**0.18**	3.06
		NNMIS	**38.4**	35.0	32.3	11.4	26.0	3.14	**3.28**	1.97	1.92	2.54	5.45	4.788	2.78	0.22	2.96
100	5	PMM	71.8	49.8	43.5	12.2	61.6	4.496	3.24	2.83	3.09	3.68	4.822	3.294	1.952	0.18	1.68
		NNMIS	**72.9**	49.6	44.2	12.5	61.8	**4.61**	3.23	3.00	3.12	3.78	4.651	3.074	1.833	**0.16**	1.37
	35	PMM	66.1	49.4	43.5	12.3	59.8	**4.38**	2.68	1.88	2.38	3.72	5.42	4.26	2.66	0.45	1.42
		NNMIS	**67.6**	48.8	44.9	12.8	59.2	4.44	2.69	1.94	2.40	3.69	5.32	4.06	2.48	**0.41**	1.44
200	5	PMM	82.4	69.2	68.1	14.4	78.1	4.70	3.697	2.925	3.195	4.46	3.692	2.924	1.669	0.25	1.24
		NNMIS	**83.1**	70.6	67.5	15.0	78.6	4.69	3.65	3.028	3.21	4.49	3.581	2.799	1.685	**0.24**	1.25
	35	PMM	80.9	68.4	64.6	14.8	77.8	4.36	3.114	2.728	3.02	4.32	3.558	2.806	1.552	0.38	1.28
		NNMIS	**81.8**	69.6	64.1	15.1	78.4	**4.38**	2.968	2.716	2.99	4.39	3.52	2.764	1.558	**0.36**	1.23

Table 3: Simulation Result for Design 3: Out of 6 active variables, 3 are fully observed and 3 have missing values; *TMIR* closer to 100 is better; $AVG_C IR$ closer to 6 is better; *AVFP* closer to 0 is better

n	Miss %	Method	TMIR					$AVG_C IR$					AVFP				
			L_1	SCAD	MCP	LRT_C	LRT_R	L_1	SCAD	MCP	LRT_C	LRT_R	L_1	SCAD	MCP	LRT_C	LRT_R
50	5	PMM	51.6	45.2	38.8	14.0	47.7	**4.04**	3.43	2.61	1.04	2.58	2.55	3.52	2.40	**0.11**	2.04
		NNMIS	**52.5**	46.2	37.8	14.2	48.8	3.99	3.32	2.5	1.09	2.56	2.58	3.59	2.46	0.12	2.05
	35	PMM	45.2	40.0	35.2	13.8	46.1	**3.973**	3.237	2.019	2.56	2.618	5.949	4.974	3.474	0.14	2.01
		NNMIS	**46.8**	39.6	35.8	13.9	46.6	3.797	3.174	2.162	2.96	2.62	6.161	5.373	4.127	**0.12**	2.03
100	5	PMM	76.8	53.4	52.8	14.2	58.9	5.02	3.898	3.312	3.22	3.56	4.273	2.931	1.796	**0.18**	1.58
		NNMIS	**77.2**	53.1	52.6	14.8	59.1	**5.11**	3.88	3.36	3.18	2.87	4.99	3.657	2.865	0.19	1.61
	35	PMM	68.5	54.8	50.2	14.1	58.5	4.832	3.845	3.293	3.26	3.69	4.611	3.929	2.137	0.20	1.63
		NNMIS	**69.1**	54.6	51.7	14.3	58.9	**4.903**	3.767	3.236	3.25	3.684	4.568	3.941	2.187	**0.19**	1.64
200	5	PMM	87.6	75.6	75.9	14.9	80.9	5.31	4.664	4.904	3.42	5.07	4.022	3.941	3.389	0.24	1.19
		NNMIS	88.8	74.9	75.4	15.2	81.1	5.416	4.597	4.818	3.28	5.68	4.006	4.101	3.898	**0.22**	1.16
200	35	PMM	80.9	73.4	72.6	15.1	78.9	**5.47**	4.32	4.71	3.30	5.19	4.438	4.091	3.651	0.26	1.19
		NNMIS	**81.6**	72.8	72.1	15.2	79.1	5.46	4.424	4.795	3.29	5.17	4.661	4.208	4.549	**0.24**	1.14

Table 4: Estimates from grouped variable selection methods with variables identified by LRTs for Overall-survival

Variables	L_1	SCAD	MCP	LRT_c	LRT_R
Age	0.0055	0	0	×	×
PVE	0	-0.671	-0.898	×	×
TOTAL number of liver Metastases	-0.00054	-0.0024	-0.0025	√	√
At least 1 cardiovascular disease	0.02204	0	0.437	×	×
Diabetes	0	-0.0025	0	×	×
Bilateral Disease	0	-0.022	-0.023	×	×
Number of NeoAdj. chemo Cycles	0	-0.022	-0.022	×	×
Extrahepatic metastasis	0	-0.077	0	×	×
Blood Losses	0	0.0006	-0.0007	×	√
Extended-Right-Resection	0	0	0	×	×
operative-time	0	0	0	√	√