Substantive-Model-Compatible Multiple Imputation for Cox Regression with a Diverging Number of Covariates

Zhilin Zhang
Department of Biostatistics, University of Michigan, Ann Arbor

Yi Li *
Department of Biostatistics, University of Michigan, Ann Arbor

Abstract

Modern biomedical survival studies with high-dimensional genomic and clinical predictors are challenged by missing covariates. Existing methods conduct inference through penalization and debiasing when the number of covariates diverges with sample size, but they are typically developed with fully observed covariates. Conversely, substantive-model-compatible multiple imputation methods, particularly substantive-model-compatible fully conditional specification (SMC-FCS), provide principled handling of missing covariates while preserving compatibility with the Cox model, yet current methodology and theory remain largely restricted to fixed-dimensional settings. To address these limitations, we propose a semiparametric multiple imputation framework for inference in Cox regression with missing covariates of a diverging dimension. Missing covariates are imputed through a high-dimensional SMC-FCS procedure driven by Cox-model likelihood contributions, with rejection sampling used to enforce substantive-model compatibility and ridge-regularized posterior draws used to stabilize the imputation models. The algorithm stabilizes the Cox estimator through an imputation-regularized optimization iteration and then generates multiply imputed datasets from a stabilized chain. Inference for low-dimensional linear functionals or contrasts, $c^\top \boldsymbol{\beta}$, is obtained by combining debiased estimators and within-imputation variance estimates through Rubin's rules. We establish consistency and asymptotic normality of the resulting pooled estimator under a diverging-dimensional regime. Simulation studies demonstrate favorable finite-sample performance, and an application to the Boston Lung Cancer Survival Cohort illustrates the practical utility of the proposed method for high-dimensional survival studies with incomplete covariates.

Keywords: Cox proportional hazards model, Statistical inference, Multiple imputation, Debiased lasso
Mathematics Subject Classification (2020): 62N02, 62J07, 62F12, 62P10.

1 Introduction

As a motivating example, we consider the Harvard School of Public Health (HSPH) subgroup of the Boston Lung Cancer Survival Cohort (BLCS), a large prospective study of lung cancer

*Corresponding author: yili@umich.edu

survivors investigating cancer genetics. The analytical population consists of 984 European-ancestry patients with pathologically confirmed non-small cell lung cancer (NSCLC), germline genotype measurements, and survival information. The analysis includes 53 single nucleotide polymorphisms (SNPs) and 10 clinical covariates, with the goal of identifying factors associated with overall survival. Although the overall missing-cell proportion is only 1.17%, and no variable has more than 5% missingness, the accumulated missing-data pattern is substantial: complete-case analysis would retain fewer than 441 patients, eliminating more than half of the cohort. More broadly, modern biomedical survival studies increasingly combine censored outcomes with high-dimensional molecular and clinical predictors, requiring methods that simultaneously address regularized estimation, incomplete covariates, and valid statistical inference.

Substantial progress has been made for Cox regression (Cox, 1972) with a diverging number of covariates ($p \to \infty$ with $p < n$). In these settings, the standard partial likelihood estimator becomes unstable, motivating penalized methods such as the LASSO (Tibshirani, 1997) and SCAD (Fan and Li, 2002) for estimation and variable selection. These methods introduced shrinkage bias and debiasing or desparsification methods were developed to remove biases due to penalization and enable confidence intervals and hypothesis testing for low-dimensional regression components (Zhang and Zhang, 2014; van de Geer et al., 2014). For survival outcomes, Xia et al. (2023) established an inferential method for Cox models with diverging covariate dimension. Unlike earlier debiased methods (Fang et al., 2017; Kong et al., 2021), which relied on sparsity assumptions on the inverse information matrix, their projection-based approach remains valid under strong covariate dependence and dense information structures. These methods are typically developed under the idealized assumption that all covariates are fully observed. As seen in our motivating example, missingness is common in clinical and epidemiologic studies. Ad hoc approaches such as complete-case analysis or single imputation can induce bias, reduce efficiency, and fail to account for uncertainty in the missing values (Little and Rubin, 2019).

A parallel literature addresses missing covariates in survival analysis through Multiple Imputation (MI) (Rubin, 1987). Standard MI approaches, including Fully Conditional Specification (FCS) (Van Buuren et al., 2006; White et al., 2011), provide flexibility for mixed data types but often fail to preserve compatibility with the substantive Cox model. This incompatibility, or lack of congeniality (Meng, 1994), biases hazard ratio estimation (White and Royston, 2009). To address this issue, Bartlett et al. (2015), Keogh and Morris (2018), and Keogh et al. (2018) proposed Substantive-Model-Compatible Fully Conditional Specification (SMC-FCS), which explicitly enforces compatibility by constructing imputations consistent with the Cox model, providing a principled way for preserving valid Cox regression inference.

The existing SMC-FCS methodology and theory are largely restricted to fixed-dimensional settings. The imputation models, computational procedures, and theoretical guarantees underlying SMC-FCS do not directly extend to regimes where the number of covariates diverges with sample size. In particular, Rubin's rules rely on classical large-sample approximations that may break down in these settings, and the interaction between imputation uncertainty and regularization remains poorly understood.

Under the diverging settings, some work has estimated models with a diverging number of miss-

ing covariates. Wang et al. (2015) proposed an EM algorithm with sparsity-enforcing truncation steps and established convergence rates together with asymptotic normality for low-dimensional components. Liang et al. (2018) introduced the Imputation-Regularized Optimization (IRO) algorithm, which alternates between imputation and regularized optimization, and established consistency of the averaged estimator under general missingness mechanisms. However, these works may not apply to Cox regression.

To address this gap, we propose a semiparametric multiple imputation method for inference in Cox regression models with a diverging number of covariates subject to missingness. The proposed method integrates Cox regression with substantive-model-compatible multiple imputation through a debiased LASSO estimator, where missing covariates are imputed using an SMC-FCS procedure adapted to diverging settings. Imputation is driven by Cox-model likelihood contributions, with rejection sampling used to enforce substantive-model compatibility. Ridge-regularized Bayesian posterior draws stabilize the conditional imputation models when the covariate dimension is large relative to the sample size, and a two-phase algorithm first stabilizes the Cox coefficients through an IRO-type iteration before generating multiply imputed datasets from a thinned posterior chain. For low-dimensional functionals $c^\top \boldsymbol{\beta}$, inference is conducted by applying Rubin's rules to combine the debiased point estimates and within-imputation variance estimators across imputations.

The proposed method offers several advantages. First, it provides a unified approach to handling missing covariates and Cox regression within a coherent inferential framework. Second, by combining debiasing with SMC-FCS, it enables valid inference for scientifically meaningful low-dimensional targets even when the covariate dimension diverges with sample size, whereas existing methods either ignore substantive-model compatibility or fail to propagate imputation uncertainty into inference. Third, we establish consistency and asymptotic normality of the pooled estimator by integrating large-sample theory for multiple imputation with Cox inference. Finally, the method is computationally feasible and broadly applicable, as illustrated through simulation studies and an analysis of the Boston Lung Cancer Survival Cohort with missing covariates.

The paper is organized as follows. Section 2 presents the proposed substantive-model-compatible debiased lasso framework, including the iterative imputation-and-optimization procedure, stationary SMC-FCS construction, debiased Cox lasso estimation, and Rubin pooling strategy. Section 3 establishes the main theoretical results. Sections 4 and 5 present simulation studies and a lung cancer survival application, respectively. Section 6 concludes with discussion and future directions. All proofs are given in the Appendix.

2 Methodology

2.1 Overview of the proposed SMC-DBL framework

We propose substantive-model-compatible debiased lasso (SMC-DBL) that integrates substantive-model-compatible multiple imputation, penalized Cox regression, and debiased inference. The procedure iteratively updates the imputations and substantive Cox model until approximate

compatibility is achieved. Specifically, missing covariates are updated through an SMC-FCS mechanism anchored to the Cox likelihood, while the regression coefficient and baseline cumulative hazard are repeatedly re-estimated from the evolving completed data. After convergence and burn-in, independently generated completed datasets are analyzed using the debiased Cox lasso (Xia et al., 2023), and the resulting estimators are combined using Rubin's rules. Additional implementation details are summarized in Algorithm 1.

2.2 Setup and Notation

Consider a clinical study with n independent subjects, where the ith subject contributes $(Y_i, \Delta_i, X_i, R_i)$ for $i = 1, \ldots, n$. The observed time is $Y_i = \min(T_i, C_i)$, where T_i denotes the failure time and C_i an independent right-censoring time, with event indicator $\Delta_i = I(T_i \le C_i)$. The covariate vector $X_i \in \mathbb{R}^p$ may be partially observed, with $R_i = (R_{i1}, \ldots, R_{ip}) \in \{0,1\}^p$ denoting the missingness pattern, where $R_{ij} = 1$ if X_{ij} is observed and 0 otherwise. We assume that T_i follows the Cox proportional hazards model

$$\lambda(t \mid X_i) = \lambda_0(t) \exp(X_i^{\mathsf{T}} \beta^0), \qquad t \ge 0,$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard and $\beta^0 \in \mathbb{R}^p$ is the regression parameter of interest. Also denote by $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ the cumulative baseline hazard. We consider a regime with $p < n$, where p is allowed to diverge with n, and denote the sparsity level by $s_0 = \|\beta^0\|_0$.

For each subject, we partition the covariates into observed and missing components as $X_i = (X_i^{\text{obs}}, X_i^{\text{mis}})$, where $X_i^{\text{obs}} = \{X_{ij} : R_{ij} = 1\}$ and $X_i^{\text{mis}} = \{X_{ij} : R_{ij} = 0\}$. Let $\mathcal{M} = \{j : \sum_{i=1}^n (1 - R_{ij}) > 0\}$ denote the set of partially observed variables. Missingness is assumed to be missing at random (MAR), namely $P(R_i \mid X_i, Y_i, \Delta_i) = P(R_i \mid X_i^{\text{obs}}, Y_i, \Delta_i)$, so that the missingness mechanism is ignorable for likelihood-based inference.

We refer to a *completed covariate matrix* as $X = (X_1, \ldots, X_n)^{\mathsf{T}} \in \mathbb{R}^{n \times p}$. Let $\ell_n(\beta; X)$ denote the negative Cox partial log-likelihood normalized by n, i.e.,

$$\ell_n(\beta; X) = -\frac{1}{n} \sum_{i=1}^n \Delta_i \left[X_i^\top \beta - \log \left\{ \sum_{j: Y_j \ge Y_i} \exp(X_j^\top \beta) \right\} \right],$$

the score function and the information matrix are respectively denoted by

$$\dot{\ell}_n(\beta; X) = -\frac{1}{n} \sum_{i=1}^n \left\{ X_i - \frac{\hat{\mu}_1(Y_i; \beta)}{\hat{\mu}_0(Y_i; \beta)} \right\} \Delta_i, \qquad \ddot{\ell}_n(\beta; X) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\hat{\mu}_2(Y_i; \beta)}{\hat{\mu}_0(Y_i; \beta)} - \left[\frac{\hat{\mu}_1(Y_i; \beta)}{\hat{\mu}_0(Y_i; \beta)} \right]^{\otimes 2} \right\} \Delta_i, \tag{1}$$

where

$$\hat{\mu}_r(t; \beta) = n^{-1} \sum_{j=1}^n I(Y_j \ge t) X_j^{\otimes r} \exp(X_j^T \beta), \qquad r = 0, 1, 2.$$

Define the weighted average covariate vector

$$\widehat{\eta}_n(t; \beta, X) = \frac{\hat{\mu}_1(t; \beta)}{\hat{\mu}_0(t; \beta)} = \frac{\sum_{j=1}^n X_j \exp(X_j^{\mathsf{T}} \beta) I(Y_j \ge t)}{\sum_{j=1}^n \exp(X_j^{\mathsf{T}} \beta) I(Y_j \ge t)}.$$

2.3 Compatibility and standard SMC-FCS

Let $f_{\mathrm{Cox}}(Y_i, \Delta_i \mid X_i; \beta, \Lambda_0)$ denote the *substantive* survival model. A collection of imputation models $\{g_j(x_j \mid X_{i,-j}, Y_i, \Delta_i)\}_{j \in \mathcal{M}}$ is said to be *compatible* with the substantive model if there exists a joint covariate distribution h on $\mathbb{R}^p$ such that, for every $j \in \mathcal{M}$,

$$g_j(x_j \mid X_{i,-j}, Y_i, \Delta_i) \;\propto\; h(x_j \mid X_{i,-j})\, f_{\mathrm{Cox}}\{Y_i, \Delta_i \mid X_i(x_j); \beta, \Lambda_0\}, \tag{2}$$

where $X_i(x_j)$ denotes X_i with its jth entry replaced by x_j. Standard FCS (Van Buuren et al., 2006; White et al., 2011) specifies each g_j directly, typically through a generalized linear model conditional on $(X_{i,-j}, Y_i, \Delta_i)$. Such conditionals are not, in general, compatible with a semi-parametric proportional-hazards substantive model. SMC-FCS (Bartlett et al., 2015) instead constructs the imputation conditional to have the form in (2). Specifically, it specifies a working covariate model $h(x_j \mid X_{i,-j}; \alpha_j)$ and samples from a density proportional to

$$h(x_j \mid X_{i,-j}; \alpha_j)\, f_{\mathrm{Cox}}\{Y_i, \Delta_i \mid X_i(x_j); \beta, \Lambda_0\}.$$

For example, rejection sampling may be implemented by proposing $x^\star \sim h(\cdot \mid X_{i,-j}; \widehat{\alpha}_j)$ and accepting it with probability

$$\frac{f_{\mathrm{Cox}}\{Y_i, \Delta_i \mid X_i(x^\star); \beta, \widehat{\Lambda}_0\}}{\sup_x f_{\mathrm{Cox}}\{Y_i, \Delta_i \mid X_i(x); \beta, \widehat{\Lambda}_0\}}.$$

Cycling over $j \in \mathcal{M}$ generates imputations whose conditional distributions incorporate the substantive-model likelihood by construction. When the working covariate conditionals correspond to a common joint distribution h, these conditionals satisfy (2) with respect to the currently imposed (β, Λ_0).

Our proposal preserves this substantive-model-compatible construction while addressing the instability arising when p diverges. The working covariate regressions employ ridge-regularized posterior or Laplace draws, a Metropolis–Hastings implementation replaces direct rejection sampling, and outer IRO updates repeatedly re-estimate (β, Λ_0) from the evolving completed data. After stabilization and burn-in, independently retained completed datasets are analyzed using the debiased Cox estimator, and the resulting estimates are combined using Rubin's rules. Thus, ridge regularization modifies the working covariate models to improve numerical stability, whereas the Cox likelihood contribution preserves compatibility with the substantive survival model.

2.4 High-dimensional SMC-FCS sweeps

Fix a missing variable $j \in \mathcal{M}$ and condition on the current completed design X, the current Cox coefficient β, and the current baseline cumulative hazard estimator $\widehat{\Lambda}_0$. Let $\mathrm{obs}_j = \{i : R_{ij} = 1\}$ index the rows on which X_j is observed, set $n_{\mathrm{obs},j} = |\mathrm{obs}_j|$, write $y_{\mathrm{obs},j} = (X_{ij})_{i \in \mathrm{obs}_j}$, and let $Z_{\mathrm{obs},j}$ be the row-centred submatrix of X_{-j} restricted to obs_j. We handle continuous and discrete covariates separately.

For continuous X_j, the working model is Gaussian. With ridge multiplier λ_{ridge} chosen as in

Section 2.7, the ridge point estimate and residual variance are

$$\widehat{\phi}_j = \left(\frac{Z_{\text{obs},j}^{\mathsf{T}} Z_{\text{obs},j}}{n_{\text{obs},j}} + \lambda_{\text{ridge}} I \right)^{-1} \frac{Z_{\text{obs},j}^{\mathsf{T}} y_{\text{obs},j}}{n_{\text{obs},j}}, \qquad \widehat{\sigma}_j^2 = \frac{1}{\text{df}} \| y_{\text{obs},j} - \widehat{\alpha}_j - Z_{\text{obs},j} \widehat{\phi}_j \|_2^2,$$

with $\text{df} = \max(n_{\text{obs},j} - p, 1)$. A Bayes-type draw of the working-model parameters is then obtained via

$$\sigma_j^{2,\star} \sim \frac{\text{df}\, \widehat{\sigma}_j^2}{\chi^2_{\text{df}}}, \qquad \phi_j^\star \sim \mathcal{N}\left(\widehat{\phi}_j, \frac{\sigma_j^{2,\star}}{n_{\text{obs},j}} \left(\frac{Z_{\text{obs},j}^{\mathsf{T}} Z_{\text{obs},j}}{n_{\text{obs},j}} + \lambda_{\text{ridge}} I \right)^{-1} \right),$$

and the corresponding intercept $\alpha_j^\star$ is reconstructed from the centring constants.

For each subject i with $R_{ij} = 0$, a proposal $x^\star \sim \mathcal{N}(\alpha_j^\star + X_{i,-j}^{\mathsf{T}} \phi_j^\star, \sigma_j^{2,\star})$ is drawn and truncated to $[-K, K]$. Let x_{curr} denote the current imputed value. The proposal is accepted with probability

$$\alpha(x_{\text{curr}}, x^\star) = \min\left\{ 1, \frac{f_{\text{Cox}}(Y_i, \Delta_i \mid x^\star, X_{i,-j}; \beta, \widehat{\Lambda}_0)}{f_{\text{Cox}}(Y_i, \Delta_i \mid x_{\text{curr}}, X_{i,-j}; \beta, \widehat{\Lambda}_0)} \right\}, \tag{3}$$

where

$$f_{\text{Cox}}(Y_i, \Delta_i \mid X_i; \beta, \widehat{\Lambda}_0) = \{ \Delta \widehat{\Lambda}_0(Y_i) \exp(X_i^\top \beta) \}^{\Delta_i} \exp\left[-\widehat{\Lambda}_0(Y_i) \exp(X_i^\top \beta) \right],$$

with $\Delta \widehat{\Lambda}_0(Y_i) = \widehat{\Lambda}_0(Y_i) - \widehat{\Lambda}_0(Y_i-)$ denoting the jump of the Breslow baseline cumulative hazard estimator at the observed failure time. For censored observations, it follows that $\Delta \widehat{\Lambda}_0(Y_i) = 0$ with $\Delta_i = 0$, and we follow the convention of $0^0 = 1$.

The accept-reject rule in equation (3) is the Metropolis-Hastings analogue of the rejection-sampling acceptance ratio used by Bartlett et al. (2015), with the Gaussian working model serving as the proposal distribution.

For binary, ordinal categorical, or unordered categorical X_j, the working model is logistic, proportional-odds, or multinomial logistic, respectively, each fitted by ridge-penalized maximum likelihood using the observed rows. Bayesian draws of the working-model parameters are obtained from a Laplace approximation centred at the ridge estimator, with covariance given by the inverse regularized observed information matrix. Candidate draws $X_{ij}^\star$ for missing entries are then sampled from the corresponding conditional distribution, and the Cox-likelihood ratio in equation (3) again determines acceptance.

2.5 Substantive-model Updates

In our proposed procedure, an *inner sweep* is one full pass of the SMC-FCS chain over all incomplete covariates conditional on the current $(\beta, \widehat{\Lambda}_0)$, while an *outer sweep* is one complete IRO iteration consisting of S_{in} inner sweeps followed by updating the regression coefficient and baseline cumulative hazard estimates using the updated completed dataset. Algorithmically, the procedure uses two levels of indexing: $\ell = 0, 1, 2, \ldots$ indexes the outer IRO iterations within a chain, while $m = 1, \ldots, M$ indexes the independently initialized inferential chains (which will be later combined through Rubin pooling).

Specifically, for chain m at outer iteration ℓ, given the current model parameters $(\beta_m^{(\ell)}, \widehat{\Lambda}_{0m}^{(\ell)})$, run S_{in} inner sweeps as described in Section 2.4, and let $\widetilde{X}_m^{(\ell+1)}$ denote the completed dataset obtained after the final inner sweep. We update the estimate of the regression coefficients by

$$\beta_m^{(\ell+1)} = \arg\min_{\beta} \left\{ \ell_n(\beta; \widetilde{X}_m^{(\ell+1)}) + \lambda_n \|\beta\|_1 \right\},$$

and the corresponding Breslow baseline cumulative hazard estimator is updated by

$$\widehat{\Lambda}_{0m}^{(\ell+1)}(t) = \sum_{r: Y_r \le t, \Delta_r = 1} \left[\sum_{i: Y_i \ge Y_r} \exp\left\{ (\widetilde{X}_{im}^{(\ell+1)})^\top \beta_m^{(\ell+1)} \right\} \right]^{-1}.$$

The outer IRO iterations continue until

$$\|\beta_m^{(\ell+1)} - \beta_m^{(\ell)}\|_1 + \|\widehat{\Lambda}_{0m}^{(\ell+1)} - \widehat{\Lambda}_{0m}^{(\ell)}\|_\infty < \varepsilon,$$

where $\varepsilon > 0$ is a pre-specified tolerance, or the maximum number of iteration steps is reached. At the final iteration step, set $(\beta_m^\star, \widehat{\Lambda}_{0m}^\star, \widetilde{X}_m^\star) = (\beta_m^{(\ell+1)}, \widehat{\Lambda}_{0m}^{(\ell+1)}, \widetilde{X}_m^{(\ell+1)})$.

2.6 Debiased Cox inference on the retained completed datasets

We first recall the fully observed construction in the notation needed below. When X is fully observed, the Cox lasso estimator is

$$\widehat{\beta}_{L_1} = \arg\min_{\beta \in \mathbb{R}^p} \{ \ell_n(\beta; X) + \lambda_n \|\beta\|_1 \}.$$

Because the penalty induces shrinkage bias, Xia et al. (2023) use the correction

$$\widehat{b} = \widehat{\beta}_{L_1} - \widehat{\Theta}\, \dot{\ell}_n(\widehat{\beta}_{L_1}; X),$$

They estimate each row of $\widehat{\Theta}$ by solving

$$\widehat{\Theta}_{j\cdot} = \arg\min_{m \in \mathbb{R}^p} m^\mathsf{T} \widehat{\Sigma}_{\widehat{\beta}_{L_1}} m \quad \text{subject to} \quad \|\widehat{\Sigma}_{\widehat{\beta}_{L_1}} m - e_j\|_\infty \le \gamma_n, \qquad j = 1, \ldots, p, \tag{4}$$

where e_j is the jth canonical basis vector and

$$\widehat{\Sigma}_{\widehat{\beta}_{L_1}} = \frac{1}{n} \sum_{i=1}^n \Delta_i \{ X_i - \widehat{\eta}_n(Y_i; \widehat{\beta}_{L_1}, X) \}^{\otimes 2}. \tag{5}$$

Here, they advocate using $\widehat{\Sigma}_{\widehat{\beta}_{L_1}}$ in place of $\ddot{\ell}_n(\widehat{\beta}_{L_1})$, as defined in (1), for theoretical convenience. Since $\left\| \widehat{\Sigma}_{\widehat{\beta}} - \ddot{\ell}_n(\widehat{\beta}) \right\|_\infty = o_p(1)$, the resulting debiased Lasso estimators are asymptotically equivalent, and their numerical difference is negligible.

With missing covariates, we apply this complete-data construction within each retained imputation before pooling. Starting from $\widetilde{X}_m^\star$, run T_0 additional SMC-FCS sweeps conditional on fixed $(\beta_m^\star, \widehat{\Lambda}_{0m}^\star)$ and ridge level $\widehat{\lambda}_{\text{ridge}}$. We retain the terminal completed dataset as $\widetilde{X}^{(m)}$, and

fit the Cox lasso on $\widetilde{X}^{(m)}$:

$$\widehat{\beta}_{L_1}^{(m)} = \arg\min_{\beta}\{\ell_n(\beta; \widetilde{X}^{(m)}) + \widehat{\lambda}_n\|\beta\|_1\}.$$

Define

$$\widehat{\Sigma}^{(m)} = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \left\{\widetilde{X}_i^{(m)} - \widehat{\eta}_n\Big(Y_i; \widehat{\beta}_{L_1}^{(m)}, \widetilde{X}^{(m)}\Big)\right\}^{\otimes 2}$$

analogous to (5). An approximate inverse information matrix $\widehat{\Theta}^{(m)}$ is constructed row by row by solving the nodewise quadratic programme

$$\widehat{\Theta}_k^{(m)} = \arg\min_{u\in\mathbb{R}^p} u^\top \widehat{\Sigma}^{(m)} u \quad \text{subject to} \quad \|\widehat{\Sigma}^{(m)}u - e_k\|_\infty \le \gamma_n, \qquad k = 1, \ldots, p, \tag{6}$$

where e_k is the kth canonical basis vector. In practice, equation (6) is solved on the positive-eigenvalue subspace of $\widehat{\Sigma}^{(m)}$, with a ridge fallback used when infeasibility is encountered.
The per-chain debiased estimator is

$$\widehat{\beta}_{\mathrm{db}}^{(m)} = \widehat{\beta}_{L_1}^{(m)} - \widehat{\Theta}^{(m)}\dot{\ell}_n^{(m)}(\widehat{\beta}_{L_1}^{(m)}).$$

Pooling across the M retained chains follows Rubin's rules. Define

$$\bar{\beta}_M = \frac{1}{M}\sum_{m=1}^{M}\widehat{\beta}_{\mathrm{db}}^{(m)}, \qquad \widehat{V}_W = \frac{1}{M}\sum_{m=1}^{M}\frac{\widehat{\Theta}^{(m)}}{n},$$

$$\widehat{V}_B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\beta}_{\mathrm{db}}^{(m)} - \bar{\beta}_M\right)\left(\widehat{\beta}_{\mathrm{db}}^{(m)} - \bar{\beta}_M\right)^\top, \qquad \widehat{V}_{\mathrm{total}} = \widehat{V}_W + \left(1 + \frac{1}{M}\right)\widehat{V}_B.$$

The resulting $(1-\alpha)$ confidence interval for the kth component is

$$\bar{\beta}_{M,k} \pm z_{\alpha/2}\sqrt{\widehat{V}_{\mathrm{total},kk}}, \qquad k = 1, \ldots, p,$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ standard normal quantile.

The standard normal reference, rather than a Barnard–Rubin t approximation, is justified by the asymptotic regime considered in Theorem 4. Inference for a fixed-dimensional linear contrast $c^\top\beta^0$ proceeds by replacing $\bar{\beta}_M$ with $c^\top\bar{\beta}_M$ and $\widehat{V}_{\mathrm{total}}$ with $c^\top\widehat{V}_{\mathrm{total}}c$ and applying the corresponding normal approximation; see Theorem 4.

2.7 Tuning of the hyper-parameters

Two hyper-parameters enter the procedure. The Cox lasso penalty λ_n is selected by five-fold cross-validation on the partial likelihood at each Cox fit, which delivers the rate $\lambda_n \asymp \sqrt{\log p/n}$ required by Huang et al. (2013). The ridge penalty is fixed at $\lambda_{\mathrm{ridge}} = \lambda_n^2(p-1)$, a scaling that aligns with the ridge prediction consistency requirement of Assumption 8 and that admits the closed-form bias-variance decomposition used in Section 3. The debiased-lasso constraint radius is set to $\gamma_n = a\sqrt{\log p/n}$ with a fixed multiplier $a = 0.5$, an interior value located in a three-point sensitivity analysis over $(a = 0.25, 0.5, 1.0)$. Empirical exploration of cross-validation

surfaces for a on thresholded debiased estimates produced essentially flat objectives across the interior of any reasonable grid, which is consistent with the order-one scaling implied by Xia et al. (2023).

Algorithm 1 SMC-DBL for Cox Regression with Iterated Baseline Hazard

Require: Observed data $\{(Y_i, \Delta_i, X_i^{\mathrm{obs}}, R_i)\}_{i=1}^n$; number of imputations M; inner sweeps S_{in}; burn-in length T_0; outer tolerance ε; CV folds K_{cv}; nominal ridge level $\lambda_{\mathrm{ridge}}^{(0)}$.

Ensure: Pooled debiased estimator $\bar{\beta}_M$, variance estimator $\widehat{V}_{\mathrm{total}}$, and confidence intervals.

Phase 0: Tuning.

Initialise missing entries by random draws from observed values within each incomplete covariate.

Run a preliminary IRO loop with inner SMC-FCS sweeps, iteratively updating both the Cox lasso coefficient β and the Breslow baseline cumulative hazard $\widehat{\Lambda}_0$, until the joint outer iterates converge. Denote the resulting completed data by $\widetilde{X}^{(0)}$.

Choose $\widehat{\lambda}_n$ by K_{cv}-fold cross-validation of the Cox partial likelihood on $\widetilde{X}^{(0)}$.

Choose $\widehat{b}$ by cross-validation for the working conditional models of the incomplete covariates using only rows where the target covariate is observed.

Set $\widehat{\lambda}_{\mathrm{ridge}} = \widehat{b}\, p \log(p)/n$ and $\gamma_n = a\sqrt{\log p/n}$.

Discard the preliminary completed dataset $\widetilde{X}^{(0)}$.

Phases 1–3: Independent inferential chains.

for $m = 1, \ldots, M$ **do**

Phase 1.m: IRO convergence.

Initialise missing entries independently and form the completed dataset $\widetilde{X}_m^{(0)}$.

Fit the Cox lasso: $\beta_m^{(0)} = \arg\min_\beta\{\ell_n(\beta; \widetilde{X}_m^{(0)}) + \widehat{\lambda}_n\|\beta\|_1\}$.

Compute

$$\widehat{\Lambda}_{0m}^{(0)}(t) = \sum_{r:Y_r \le t, \Delta_r = 1} \Big[\sum_{i:Y_i \ge Y_r} \exp\{(\beta_m^{(0)})^\top \widetilde{X}_{im}^{(0)}\} \Big]^{-1}.$$

for $\ell = 0, 1, 2, \ldots$ **do**

I-step. Conditional on $(\beta_m^{(\ell)}, \widehat{\Lambda}_{0m}^{(\ell)})$, run S_{in} SMC-FCS sweeps: in each sweep, visit every incomplete X_j once and update its missing entries from a conditional distribution proportional to $f_j(x_j \mid X_{-j}; \widehat{\alpha}_j, \widehat{\lambda}_{\mathrm{ridge}})\, f_{\mathrm{Cox}}\{Y, \Delta \mid X_j = x_j, X_{-j}; \beta_m^{(\ell)}, \widehat{\Lambda}_{0m}^{(\ell)}\}$. Retain the terminal completed dataset and denote it by $\widetilde{X}_m^{(\ell+1)}$.

RO-step. Refit the Cox lasso: $\beta_m^{(\ell+1)} = \arg\min_\beta\{\ell_n(\beta; \widetilde{X}_m^{(\ell+1)}) + \widehat{\lambda}_n\|\beta\|_1\}$.

Update

$$\widehat{\Lambda}_{0m}^{(\ell+1)}(t) = \sum_{r:Y_r \le t, \Delta_r = 1} \Big[\sum_{i:Y_i \ge Y_r} \exp\{(\beta_m^{(\ell+1)})^\top \widetilde{X}_{im}^{(\ell+1)}\} \Big]^{-1}.$$

Stop if $\|\beta_m^{(\ell+1)} - \beta_m^{(\ell)}\|_1 + \sup_{t \le \tau} |\widehat{\Lambda}_{0m}^{(\ell+1)}(t) - \widehat{\Lambda}_{0m}^{(\ell)}(t)| < \varepsilon$.

end for

Set $(\beta_m^\star, \widehat{\Lambda}_{0m}^\star, \widetilde{X}_m^\star) = (\beta_m^{(\ell+1)}, \widehat{\Lambda}_{0m}^{(\ell+1)}, \widetilde{X}_m^{(\ell+1)})$.

Phase 2.m: Burn-in.

Starting from $\widetilde{X}_m^\star$, run T_0 additional SMC-FCS sweeps conditional on fixed $(\beta_m^\star, \widehat{\Lambda}_{0m}^\star)$ and ridge level $\widehat{\lambda}_{\text{ridge}}$.

Retain the terminal completed dataset as $\widetilde{X}^{(m)}$.

Phase 3.m: Debiased lasso inference.

Fit the Cox lasso on $\widetilde{X}^{(m)}$: $\widehat{\beta}_{L_1}^{(m)} = \arg\min_\beta\{\ell_n(\beta; \widetilde{X}^{(m)}) + \widehat{\lambda}_n\|\beta\|_1\}$.

Solve the nodewise quadratic programs with constraint radius γ_n to obtain $\widehat{\Theta}^{(m)}$.

Compute the debiased estimator $\widehat{\beta}_{\text{db}}^{(m)} = \widehat{\beta}_{L_1}^{(m)} - \widehat{\Theta}^{(m)}\dot{\ell}_n(\widehat{\beta}_{L_1}^{(m)}; \widetilde{X}^{(m)})$.

Estimate the corresponding within-imputation covariance matrix $\widehat{V}^{(m)}$.

end for

Phase 4: Rubin pooling.

Compute the pooled estimator $\bar{\beta}_M = M^{-1}\sum_{m=1}^M \widehat{\beta}_{\text{db}}^{(m)}$.

Compute the within-imputation covariance $\widehat{V}_W = M^{-1}\sum_{m=1}^M \widehat{V}^{(m)}$.

Compute the between-imputation covariance $\widehat{V}_B = (M-1)^{-1}\sum_{m=1}^M(\widehat{\beta}_{\text{db}}^{(m)} - \bar{\beta}_M)(\widehat{\beta}_{\text{db}}^{(m)} - \bar{\beta}_M)^\top$.

Compute the Rubin variance estimator $\widehat{V}_{\text{total}} = \widehat{V}_W + (1 + M^{-1})\widehat{V}_B$.

Construct confidence intervals using the diagonal entries of $\widehat{V}_{\text{total}}$.

3 Theoretical Results

We begin with the assumptions required for the asymptotic analysis.

Assumption 1. *There exists $K < \infty$ such that $\|X_i\|_\infty \le K$ almost surely for all $i = 1, \dots, n$.*

Assumption 2. *$|X_i^\mathsf{T}\beta^0| \le K_1 < \infty$ uniformly for all $i = 1, \dots, n$ almost surely.*

Assumption 3. *The maximum follow-up time is $0 < \tau < \infty$ with $\pi_0 = \mathbb{P}(Y \ge \tau) > 0$.*

Assumption 4. *The true baseline cumulative hazard Λ_0^0 is continuous, nondecreasing on $[0, \tau]$, and satisfies $\Lambda_0^0(\tau) \le C_\Lambda < \infty$.*

Assumption 5. *The population information matrix $\Sigma_{\beta^0} = \mathbb{E}[\{X_i - \eta_0(Y_i; \beta^0)\}^{\otimes 2}\Delta_i]$ satisfies $0 < \zeta_{\min} \le \zeta_{\min}(\Sigma_{\beta^0}) \le \zeta_{\max}(\Sigma_{\beta^0}) \le \zeta_{\max} < \infty$.*

Assumption 6. *For each partially observed covariate X_j, let $q_{ij}(\cdot)$ denote the working conditional imputation distribution for subject i, chosen according to the data type of X_j:*

1. *If X_j is continuous, $q_{ij}(\cdot)$ is a Gaussian density with ridge-estimated mean $\hat{\mu}_{ij}$ and variance $\hat{\sigma}_j^2$, truncated to $[-K, K]$ under Assumption A1.*
2. *If X_j is binary with support $\{0, 1\}$, $q_{ij}(\cdot)$ is a Bernoulli mass function with success probability $\hat{\pi}_{ij}$, estimated via ridge-penalized logistic regression.*
3. *If X_j is an additively coded SNP with support $\{0, 1, 2\}$, $q_{ij}(\cdot)$ is an ordinal logistic mass function with probabilities $\hat{\pi}_{ij}^{(0)}, \hat{\pi}_{ij}^{(1)}, \hat{\pi}_{ij}^{(2)}$, estimated via ridge-penalized proportional-odds regression.*

4. *If X_j is categorical with $k_j \geq 3$ unordered levels, $q_{ij}(\cdot)$ is a categorical mass function with probabilities $\hat{\pi}_{ij}^{(\ell)}$, estimated via ridge-penalized multinomial logistic regression.*

In all cases, there exist constants $0 < c_\theta \leq C_\theta < \infty$, independent of (n, p, i, j), such that $c_\theta \leq q_{ij}(x) \leq C_\theta$ for all x in the support of X_j. For discrete working models, these bounds are enforced through numerical clipping of fitted probabilities, namely $\hat{\pi}_{ij}^{(\ell)} \in [\epsilon, 1-\epsilon]$ for some fixed $\epsilon > 0$.

Assumption 7. *(a) Bounded total missingness: $d = |\{(i,j) : R_{ij} = 0\}| \leq \bar{r}\, np$ for some $\bar{r} \in (0,1)$ bounded away from 1; (b) Bounded subject-level missingness: $r_{\text{sub}} = n_{\text{mis}}/n \leq \bar{r}_{\text{sub}}$ for some $\bar{r}_{\text{sub}} \in (0,1)$; (c) The missingness mechanism is MAR.*

Assumption 8. *For each partially observed covariate X_j, let α_j^0 denote the true parameter vector of the corresponding working imputation model, where $\alpha_j^0 = \gamma_j^0$ for continuous variables, α_j^0 is the logistic regression coefficient vector for binary and categorical variables, and α_j^0 additionally includes threshold parameters for ordinal logistic models. We assume $p\|\alpha_j^0\|_2^2/\sqrt{n} = o(n^{-1/2})$.*

Assumption 9. *Let $s_0 = \|\beta^0\|_0$ and $s_* = \max_j \|\gamma_j^0\|_0$. Assume $\|\Theta_{\beta^0}\|_{1,1}^2 \log p \cdot \max(ps_0, s_*^2) = o(n)$, where $\Theta_{\beta^0} = \Sigma_{\beta^0}^{-1}$ and $\|\cdot\|_{1,1}$ is the matrix operator norm induced by the ℓ_1 vector norm.*

Assumption 10. *The number of inner SMC-FCS sweeps S_{in} satisfies $S_{\text{in}} \geq c_S n^{c_d} \log(np)$, for some $c_d = c\log(1/\varepsilon_a)$, where $\varepsilon_a = c_f c_\theta/(C_f C_\theta) > 0$ is the minorisation constant from Theorem 1, c_f, C_f are lower and upper bounds on the Cox likelihood on $[-K, K]$, and $c_S > 1$.*

Assumption 11. *The number of outer IRO iterations l at convergence satisfies $l \geq c_l \log n/\log(1/\lambda_*)$, where $\lambda_* \in (0,1)$ is the contraction constant from Assumption 12 and $c_l > 1$.*

Assumption 12. *Let $\theta = (\beta, \Lambda_0)$, and let $\mathcal{M}(\theta)$ denote the population IRO map induced by the stationary SMC-FCS law followed by the Cox lasso and Breslow updates. There exist a fixed point $\theta^* = (\beta^*, \Lambda_0^*)$, a neighbourhood $\mathcal{N}(\theta^*)$, constants $\lambda_* \in (0,1)$ and $a_n = o(1)$ such that*

$$\|\mathcal{M}(\theta) - \theta^*\|_{\mathcal{H}} \leq \lambda_* \|\theta - \theta^*\|_{\mathcal{H}} + a_n, \qquad \theta \in \mathcal{N}(\theta^*),$$

where $\|\theta - \theta^*\|_{\mathcal{H}} = \|\beta - \beta^*\|_1 + \|\Lambda_0 - \Lambda_0^*\|_\infty$.

Assumption 13. *The burn-in length T_0 per chain satisfies $T_0 \geq c_T \log(np)/\varepsilon_0$. This ensures that after T_0 iterations, each chain's terminal state is within $O(\rho^{T_0})$ total-variation distance of the stationary distribution $\nu_{\beta^\star, \Lambda_0^\star}$, where $\rho \in (0,1)$ is the geometric ergodicity rate.*

Assumptions 1–5 are standard regularity conditions for high-dimensional Cox regression (Xia et al., 2023). Assumption 6 imposes non-degeneracy of the working conditional models and yields the uniform minorisation condition required for geometric ergodicity of the Metropolis–Hastings kernel in Theorem 1, following the substantive-model-compatible FCS (Bartlett et al., 2015). Assumption 7 formalises the MAR mechanism and excludes degenerate missingness regimes. Assumptions 8–9 specify the sparsity and ridge-regularisation regime needed to control the debiasing remainder and establish asymptotic normality of the debiased Cox estimator (Xia et al., 2023). Assumptions 10–11 are algorithmic conditions ensuring sufficient inner-chain mixing and valid multiple-imputation inference (Wang and Robins, 1998; Hughes et al., 2014).

Assumption 12 is a local stochastic stability condition on the imputation-and-optimization map, controlling the asymptotic behavior of the outer IRO iterates. The assumption is supported in practice by the convexity of the Cox lasso update, ridge-stabilized working models, bounded linear predictors, and empirical stabilization of successive iterates. Assumption 13 ensures that the retained completed datasets used for Rubin pooling are approximately sampled from the stationary imputation distribution associated with the limiting substantive Cox model.

Theorem 1 (Geometric ergodicity of the inner SMC-FCS chain)**.** *Fix β with $\|\beta - \beta^0\|_1 \le r$ for some fixed $r > 0$, and let $\widehat{\Lambda}_0$ be the Breslow baseline cumulative hazard estimator computed from the current completed data and the current value of β. Fix the missingness count d. Under Assumptions 1, 2, 3, 4, 6, and 7, with probability tending to one, the inner SMC-FCS sweep kernel $P_{\beta,\widehat{\Lambda}_0}$ admits a unique stationary distribution $\nu_{\beta,\widehat{\Lambda}_0}$. Moreover, there exist constants $C_0 < \infty$ and $\rho_d \in (0,1)$, where $\rho_d = 1 - \varepsilon_a^d$ for some constant $\varepsilon_a \in (0,1)$, such that for every initial state $x_0 \in \mathcal{X}$ and every $S \ge 1$,*

$$\left\| P^S_{\beta,\widehat{\Lambda}_0}(x_0, \cdot) - \nu_{\beta,\widehat{\Lambda}_0} \right\|_{\mathrm{TV}} \le C_0 \rho_d^S.$$

Consequently, for target total-variation accuracy $\delta > 0$, it suffices that $S \ge \log(C_0/\delta)/\varepsilon_a^d$.

The results imply that the inner SMC-FCS chain converges geometrically fast to a unique stationary distribution conditional on the current substantive-model parameters. Consequently, after a logarithmic number of inner sweeps, the completed datasets used in the outer IRO updates become asymptotically insensitive to initialization, providing the key mixing condition needed for the convergence of the outer iterates established in the next theorem.

Theorem 2 (Convergence of the IRO iterates)**.** *Let $(\widetilde{\beta}^{(\ell)}, \widehat{\Lambda}_0^{(\ell)})$ denote the outer IRO iterates, where $\widetilde{\beta}^{(\ell)}$ is the Cox lasso estimator computed from the completed data after the ℓth imputation update and $\widehat{\Lambda}_0^{(\ell)}$ is the corresponding Breslow estimator. Under Assumptions 1–7, 10, 11, and Assumption 12, suppose $\sup_{n,\ell} \mathbb{E}\|\widetilde{\beta}^{(\ell)}\|_1 < \infty$. Then, for sufficiently large n and sufficiently large ℓ,*

$$\|\widetilde{\beta}^{(\ell)} - \beta^*\|_1 + \|\widehat{\Lambda}_0^{(\ell)} - \Lambda_0^*\|_\infty = o_p(1),$$

for almost every observed-data sequence, where (β^*, Λ_0^*) is the fixed point of the population IRO map.

Theorem 2 shows that the coupled imputation-and-optimization procedure converges to a stable fixed point of the population IRO map for both the regression coefficient and the baseline cumulative hazard. Combined with Theorem 1, this implies that the finite-sweep SMC-FCS updates introduce asymptotically negligible error, so the completed datasets behave asymptotically as draws from a stationary imputation mechanism compatible with the substantive Cox model. This stationary regime provides the foundation for the high-dimensional estimation and inference results developed in Theorems 3 and 4.

Theorem 3 (Cox lasso rate under IRO stationarity)**.** *Suppose the conditions of Theorem 2 hold together with Assumptions 5, 8, and 9. Assume the outer IRO iterates have reached stationarity at the fixed point (β^*, Λ_0^*). Let $(\widetilde{\beta}^{(\ell)}, \widehat{\Lambda}_0^{(\ell)})$ denote the Cox lasso estimator and Breslow baseline*

cumulative hazard estimator obtained at the ℓth outer IRO iteration from the corresponding completed dataset. Then, it holds that

$$\|\widetilde{\beta}^{(\ell)} - \beta^*\|_1 = O_p\Big(s_0\sqrt{\log p/n}\Big).$$

If additionally $\|\beta^* - \beta^0\|_1 = O\Big(s_0\sqrt{\log p/n}\Big)$ holds, then

$$\|\widetilde{\beta}^{(\ell)} - \beta^0\|_1 = O_p\Big(s_0\sqrt{\log p/n}\Big).$$

Moreover,

$$\|\widehat{\Lambda}_0^{(\ell)} - \Lambda_0^*\|_\infty = o_p(1).$$

The results suggest that, after the IRO algorithm reaches stationarity, the Cox lasso estimator retains the standard high-dimensional ℓ_1 convergence rate despite missing covariates and iterative imputation updates, while the associated Breslow baseline cumulative hazard estimator remains uniformly consistent. These rate results lead to the asymptotic normality of the debiased Rubin-pooled estimator in Theorem 4.

Theorem 4 (Asymptotic normality for linear contrasts)**.** *Suppose Assumptions 1–13 hold. Assume $\lambda_n \asymp \sqrt{\log p/n}$ and $\gamma_n \asymp \|\Theta_{\beta^0}\|_{1,1} s_0 \lambda_n$, and suppose the conclusions of Theorem 3 hold. Assume the SMC-FCS imputation procedure is asymptotically proper and congenial with the Cox substantive model in the sense of Wang and Robins Wang and Robins (1998). Let $\bar{\beta}_M = M^{-1}\sum_{m=1}^M \widehat{\beta}_{\mathrm{db}}^{(m)}$ denote the Rubin-pooled debiased estimator. Then, for any fixed loading vector $c \in \mathbb{R}^p$ satisfying $\|c\|_2 = 1$ and $\|c\|_1 \le a^* < \infty$, it holds that*

$$\frac{\sqrt{n}\, c^\top(\bar{\beta}_M - \beta^0)}{\sqrt{n\, c^\top \widehat{V}_{\mathrm{total}} c}} \xrightarrow{d} \mathcal{N}(0,1).$$

Also, it follows that

$$n\,\mathrm{Var}(c^\top(\bar{\beta}_M - \beta^0)) \to c^\top\left[V_{\mathrm{com}} + \left(1 + \frac{1}{M}\right) V_{\mathrm{mis}}\right] c,$$

where $V_{\mathrm{mis}} = V_{\mathrm{ML}} - V_{\mathrm{com}}$ and $V_{\mathrm{com}} = \Theta_{\beta^0}$, and

$$n\, c^\top \widehat{V}_{\mathrm{total}} c \xrightarrow{p} c^\top\left[V_{\mathrm{com}} + \left(1 + \frac{1}{M}\right) V_{\mathrm{mis}}\right] c,$$

so Rubin-based Studentization is asymptotically valid for every fixed $M \ge 2$.

Theorem 4 establishes asymptotically valid inference for fixed low-dimensional contrasts after debiasing and Rubin pooling across imputations. The result shows that the proposed SMC-DBL procedure correctly propagates uncertainty from missing covariates and attains the semiparametric observed-data efficiency bound when the number of imputations diverges, while remaining asymptotically valid for fixed finite M.

4 Simulations

We investigated the operating characteristics of the proposed SMC-DBL procedure through simulation. Four competitors are considered, including an oracle debiased lasso fitted to the fully observed data, the IRO procedure of Liang et al. (2018) without debiasing, the standard SMC-FCS imputation of Bartlett et al. (2015) followed by per-imputation debiased lasso and Rubin pooling, the mean-or-mode imputation followed by debiased lasso and the complete case analysis with debiased lasso. The oracle is treated as a gold standard in the sense that it has access to data the other procedures must reconstruct from observed entries.

We considered all combinations of $n \in \{500, 1000, 2000\}$ and $p \in \{20, 50, 100, 200\}$, yielding 12 settings. The number of nonzero coefficients was fixed at $s_0 = 5$, with

$$\beta^0 \;=\; (1.0,\, 1.0,\, 1.0,\, 0.5,\, 0.5,\, 0,\, \ldots,\, 0)^{\mathsf{T}}.$$

Covariates were drawn from a multivariate Gaussian distribution with mean zero and AR(1) covariance $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.5$, then truncated coordinatewise to $[-5, 5]$ to satisfy the boundedness assumption underlying the Cox-model partial-likelihood theory. Failure times were drawn from an exponential distribution with rate $\exp(X^{\mathsf{T}}\beta^0)$, and censoring times were drawn independently from an exponential distribution with rate 0.1. The marginal censoring fraction ranged from approximately 30% to 45% across settings.

Missingness was placed on coordinates $2, \ldots, \lfloor 0.2p \rfloor + 1$, so that 20% of covariates were partially missing in each setting. The missingness indicator at coordinate j followed a Bernoulli draw with probability $\text{logit}^{-1}(-1.5 + 0.8X_1)$ clipped to $[0.80, 0.95]$. The resulting mechanism is missing at random, with the fully observed first coordinate as the sole driver, matching the conditioning used in the theoretical development. The realised per-coordinate missing rate was approximately 5–20% on affected columns, with a subject-level missing rate of roughly 50% under union over coordinates and an overall cell-level rate of approximately 3–4% on the design matrix.

For each (n, p) pair, we generated $R = 100$ independent datasets and applied all five procedures. The SMC-DBL procedure used $M = 20$ inferential chains, each running an independent Phase 1 IRO convergence followed by a per-chain burn-in of $T_{\text{burn}} = \max\{20, \lceil 3\log(np)\rceil\}$ iterations, with $S_{\text{inner}} = \max\{5, \lceil \log(np)\rceil\}$ inner sweeps per outer iteration. The ridge multiplier was tuned by five-fold cross-validation on the prediction error of the first variable in $\mathcal{M}$, with the response taken from rows on which that variable is genuinely observed and the predictor matrix read from the same rows of the imputed pseudo-complete dataset produced by a single tuning Phase 1 chain run at a small nominal ridge $b_0 = 0.1$. The minimum-MSE rule was used over a candidate grid for b ranging from 0.01 to $\max(2, 20\sqrt{p/n})$ on a 15-point log-spaced grid. The QP multiplier in the debiasing step was fixed at $a = 0.5$. The converged Phase 1 state of the tuning chain was discarded, and the M inferential chains shared no Phase 1 endpoint and were initialised independently.

Three operating-characteristic figures summarise the comparison on the active set $\mathcal{S} = \{k : \beta^0_k \neq 0\}$. Per-coordinate detail across all settings is reported in Appendix Table S.2. The active set is the inferential target of practical interest, and reporting averages over $\mathcal{S}$ yields a compact

summary that nevertheless preserves the essential contrast among methods. Mean absolute bias on $\mathcal{S}$ is $\overline{|\mathrm{Bias}|}_{\mathcal{S}} = |\mathcal{S}|^{-1} \sum_{k \in \mathcal{S}} |\mathrm{Bias}_k|$, with $\mathrm{Bias}_k = R^{-1} \sum_{r=1}^{R} (\hat{\beta}_k^{(r)} - \beta_k^0)$, and the active-set averages of empirical SD, mean estimated SE, and coverage are defined analogously.

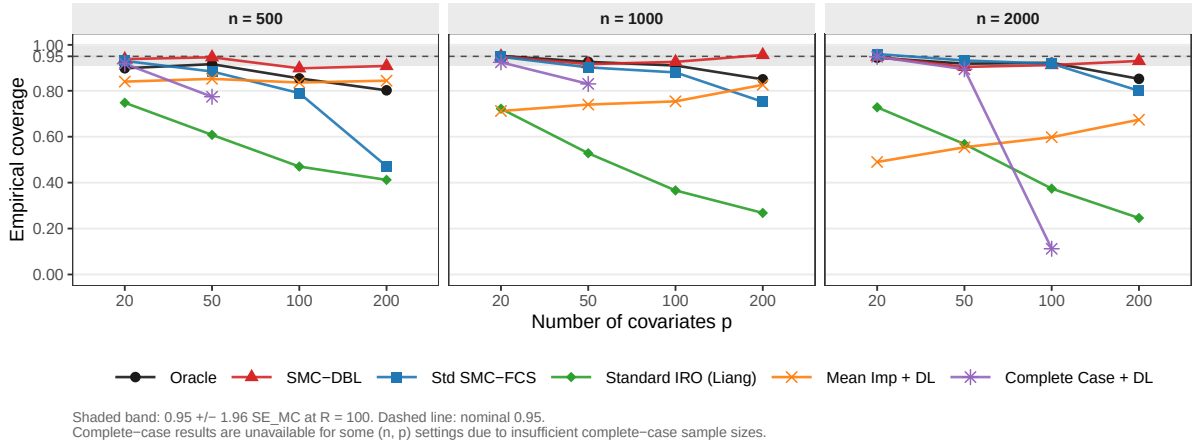


Figure 1: Empirical coverage on the active set across the twelve simulation settings.

SMC-DBL holds the nominal 0.95 level within Monte-Carlo error in eleven of the twelve settings, Std SMC-FCS achieves nominal coverage at $n = 1000$ and $n = 2000$ for $p \leq 100$ but drops to 0.47 at $(n = 500, p = 200)$ and to 0.75 at $(n = 1000, p = 200)$. The oracle, despite having access to the fully observed data, exhibits mild under-coverage (roughly 0.80–0.85) at $p = 200$ for all n, a feature attributable to the diverging-p remainder in the debiased-lasso decomposition rather than to the imputation step. Standard IRO under-covers severely and the gap widens with p: from 0.75 at $(n = 500, p = 20)$ down to 0.25–0.41 at $p = 200$, reflecting the absence of a debiasing correction. Mean Imp + DL produces low coverage that worsens with n at small p, illustrating that single-imputation procedures fail to reflect imputation uncertainty, when n grows the standard errors shrink at the $\sqrt{n}$ rate but the imputation bias does not, so the coverage gap widens. Complete Case + DL is competitive only in the lowest-dimensional settings. Its coverage is close to nominal at $p = 20$, but declines markedly at $p = 50$ and falls to 0.11 at $(n = 2000, p = 100)$. Moreover, complete-case estimates are unavailable for several larger-p settings because too few complete observations remain for model fitting.

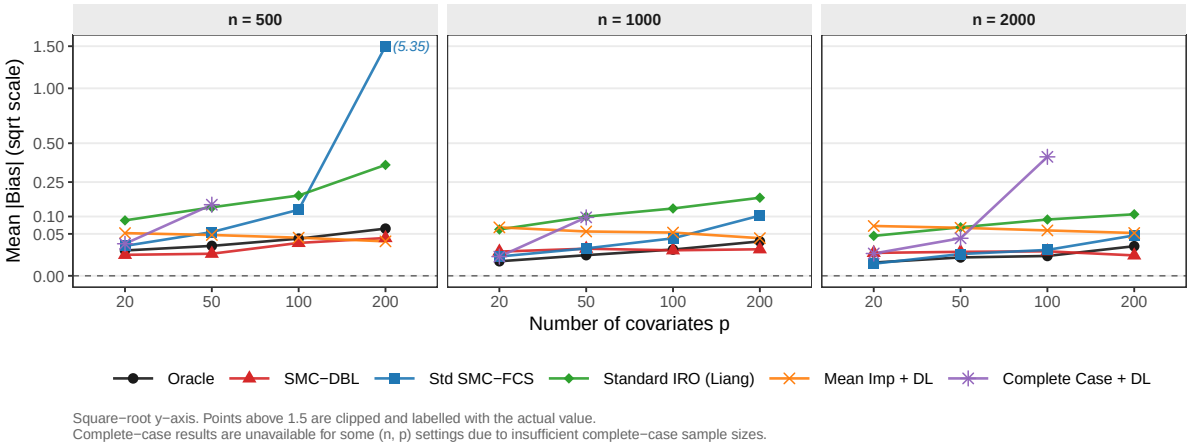


Figure 2: Mean absolute bias on the active set.

On the square-root scale, SMC-DBL tracks the oracle closely across all twelve settings, with the active-set absolute bias in the range 0.02 to 0.05. The two methods are visually indistinguishable at $n \in \{1000, 2000\}$. Std SMC-FCS tracks SMC-DBL at $p \leq 100$ but exhibits a catastrophic failure at $(n = 500, p = 200)$ where the active-set bias reaches approximately 5.5. The failure reflects the breakdown of unregularised univariate imputation in the regime $p \approx n/2$ with 20% of columns missing. Standard IRO carries a non-trivial active-set bias at every setting, in the range 0.05 to 0.35, increasing with p and decreasing with n as expected for a Cox-lasso point estimate without debiasing. Mean Imp + DL has bias comparable to the oracle for small p but drifts upward as p grows, again reflecting the loss of correlation information when missing entries are replaced by column means. Complete Case + DL also becomes unstable as the effective complete-case sample size becomes small relative to p: although its bias is small at $p = 20$, it rises to approximately 0.10–0.12 at several $p = 50$ settings and to about 0.36 at $(2000, 100)$.

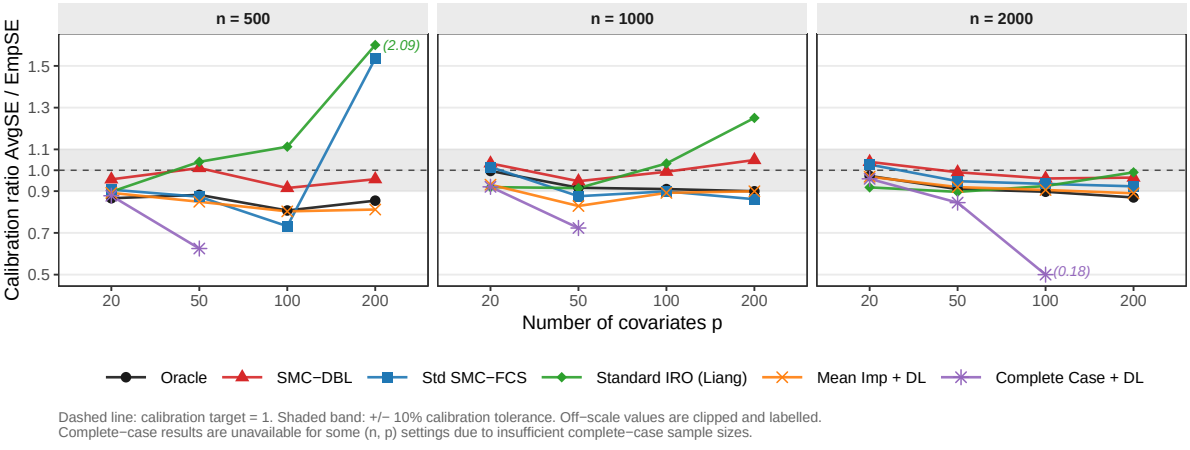


Figure 3: Variance calibration on the active set

The calibration ratio $\widehat{\text{AvgSE}}/\widehat{\text{EmpSE}}$ takes the value one when the model-based standard errors target the empirical sampling SD; values below one indicate under-estimation of sampling variability, values above one indicate conservatism. SMC-DBL falls inside the $\pm 10\%$ tolerance band in eleven of the twelve settings, with the largest deviation 0.92 occurring at $(n = 500, p = 100)$.

The ratio rises to 1.05 at $(n = 1000, p = 200)$, in line with the proper-imputation limit $V_{\mathrm{ML}} + V_{\mathrm{mis}}/M$ from Theorem 4, which inflates the within-imputation sandwich by a factor that grows with the missing-information fraction. Mean Imp + DL clusters near 0.85–0.90 for $n \leq 1000$ but moves into the calibration band at $n = 2000$, showing that single-imputation under-estimation persists at moderate n but attenuates as the sampling variance dominates the imputation variance. Std SMC-FCS is well calibrated at $p \leq 50$ but jumps to 1.55 at $(n = 500, p = 200)$, a mechanical consequence of the catastrophic bias point in Figure 2: the empirical SD on the active set is dominated by the bias spike, but the model-based sandwich does not register that spike, so the ratio is in fact misleadingly close to one only because both numerator and denominator are large rather than because the procedure is calibrated. Standard IRO produces ratios above the upper tolerance at $p \geq 100$ for $n \in \{500, 1000\}$, peaking near 1.7 at $(n = 500, p = 200)$; the plug-in Hessian-inverse used in the IRO variance estimate is over-estimating the sampling SD in regimes where shrinkage bias dominates the Monte-Carlo fluctuation. Complete Case + DL shows the strongest under-estimation as dimensionality increases: its calibration ratio declines from approximately 0.9–1.0 at $p = 20$ to roughly 0.6–0.85 at $p = 50$, and to only 0.18 at $(2000, 100)$.

5 Data Analysis

5.1 Data

The application uses a non-small cell lung cancer sub-cohort comprising 979 European-ancestry patients with pathologically confirmed disease, available germline genotype data, and complete survival information. 5 patients with missing outcomes were excluded from the analysis. Genotyping was performed on Illumina arrays. Standard sample- and variant-level quality control was applied within the cohort, including filters on sample missingness, variant call rate, minor allele frequency, and Hardy-Weinberg equilibrium. Cryptic relatedness was addressed during upstream genotype quality control. Three within-cohort ancestry principal components, computed by principal component analysis on linkage-disequilibrium-pruned genome-wide markers (PLINK v1.9, –indep-pairwise 50 5 0.2), were retained for adjustment. Overall survival was defined as the time from pathological diagnosis to death from any cause, with patients still alive at last follow-up censored at that date.

Seven clinical covariates were extracted: age at diagnosis (continuous), sex (binary), smoking status (never, former, or current), disease stage (early stages I–II vs late stages III–IV), surgical resection (yes or no), chemotherapy (yes or no), radiation therapy (yes or no), and the first three within-cohort ancestry principal components. The treatment indicators record whether each treatment was ever received, recorded at the time of data abstraction. Because treatment timing was not uniformly available, the indicators may be subject to immortal-time bias, and we therefore treat their coefficients as adjustment quantities rather than causal effect estimates.

The genetic predictors were drawn from a prespecified panel of 77 candidate single-nucleotide polymorphisms assembled from lung-function GWAS evidence, primarily from the Kachuri et al. (2020) instrument set, supplemented with representative lead variants from established

pleiotropic loci. Variants were located by chromosomal position under the GRCh37/hg19 reference build. After quality control, 51 of the 77 candidate variants were available in the cohort and entered the analysis. The remaining variants were either monomorphic, failed Hardy-Weinberg or call-rate filters, or were not present on the genotyping array. The detailed descriptive table is reported in Appendix Table S.1.

5.2 Comparisons of Results

We applied the SMC-DBL procedure described in Section 2, using a design vector that included both clinical covariates and HSPH SNP genotypes. Clinical covariates were standardised to unit variance prior to penalisation, whereas SNPs were retained on the additive 0/1/2 dosage scale. The resulting design matrix contained $p = 61$ covariates.

Table 1 compares SMC-DBL with several competing approaches considered in the simulation studies and further illustrates the advantages of the proposed method. Compared with complete-case (CC) analysis, SMC-DBL generally produced smaller standard errors and stronger statistical significance for both clinical covariates and SNP effects, reflecting improved efficiency through the use of partially observed subjects rather than discarding incomplete cases. For example, the estimated effect of late-stage disease increased from 0.772 (SE = 0.257) under CC analysis to 0.929 (SE = 0.173) under SMC-DBL. SMC-DBL and standard SMC-FCS yielded highly similar estimates, suggesting that the proposed debiased lasso method preserves the substantive-model-compatible imputation structure while extending inference to diverging-dimensional settings. In contrast, the IRO approach substantially attenuated most SNP effects toward zero, frequently producing p-values close to one. For instance, rs11022690 was statistically significant under SMC-DBL ($\hat{\beta} = -0.113$, $p = 0.022$) but was considerably attenuated under IRO ($\hat{\beta} = -0.054$, $p = 0.300$). Similar patterns were observed for rs72743477, indicating that SMC-DBL better preserves moderate prognostic signals in the presence of missingness. Although mean imputation produced estimates numerically similar to SMC-DBL for some variables, it does not properly account for uncertainty due to missing data and therefore lacks formal inferential validity. In contrast, SMC-DBL combines compatible multiple imputation, Rubin's variance decomposition, and debiased penalized estimation in a unified fashion, providing principled inference for moderately high-dimensional Cox regression with missing covariates.

Among the identified variants, rs72743477 is located within *SMAD3*, a central mediator of the TGF-β signaling pathway that regulates cell proliferation, differentiation, immune modulation, and extracellular matrix remodeling. Aberrant TGF-β/*SMAD3* signaling has been widely implicated in lung cancer progression, particularly in epithelial–mesenchymal transition (EMT), tumor invasion, metastatic dissemination, and resistance to therapy. In NSCLC, increased *SMAD3* activity has been associated with aggressive tumor phenotypes, immune evasion, fibrosis-related stromal remodeling, and poorer prognosis. Experimental studies have also suggested that *SMAD3* contributes to tumor-promoting inflammatory responses and may interact with smoking-related oxidative stress pathways, both of which are relevant to lung carcinogenesis. These biological roles support the plausibility that variation within the *SMAD3* region could influence survival outcomes through effects on tumor progression and the tumor microenvironment. In contrast, rs11022690 currently has limited functional annotation and has

not been extensively characterized in lung cancer studies. Nevertheless, the consistency of its estimated protective effects across multiple imputation-based analyses suggests a potentially stable prognostic signal rather than a spurious association driven by missing-data handling. It is possible that rs11022690 tags a nearby regulatory region or acts through linkage disequilibrium with functional variants affecting gene expression or immune-related pathways. Additional analyses integrating expression quantitative trait locus (eQTL) data, epigenomic annotations, and transcriptomic profiling may help clarify its biological relevance.

Table 1: Pooled estimates for all predictors across the five comparison procedures applied to the NSCLC cohort.

Variable	CC		SMC-DBL		Std SMC-FCS		IRO		Mean Imp	
	$\hat{\beta}$ (SE)	P-value	$\hat{\beta}$ (SE)	P-value	$\hat{\beta}$ (SE)	P-value	$\hat{\beta}$ (SE)	P-value	$\hat{\beta}$ (SE)	P-value
Clinical covariates										
Surgery (yes vs no)	-1.252 (0.179)	<0.001	-1.117 (0.123)	<0.001	-1.100 (0.124)	<0.001	-1.000 (0.116)	<0.001	-1.115 (0.123)	<0.001
Late stage (III-IV vs I-II)	0.772 (0.257)	0.003	0.929 (0.173)	<0.001	0.955 (0.176)	<0.001	0.586 (0.163)	<0.001	0.930 (0.173)	<0.001
Age (per SD)	0.222 (0.054)	<0.001	0.259 (0.038)	<0.001	0.257 (0.038)	<0.001	0.194 (0.041)	<0.001	0.258 (0.038)	<0.001
Sex (male vs female)	-0.280 (0.104)	0.007	-0.328 (0.071)	<0.001	-0.328 (0.071)	<0.001	-0.200 (0.074)	0.007	-0.326 (0.071)	<0.001
Smoking (per level)	0.205 (0.083)	0.014	0.232 (0.059)	<0.001	0.232 (0.059)	<0.001	0.122 (0.061)	0.047	0.231 (0.059)	<0.001
Radiation (yes vs no)	-0.093 (0.128)	0.471	-0.172 (0.088)	0.051	-0.175 (0.088)	0.048	0.000 (0.100)	1.000	-0.175 (0.088)	0.048
Chemotherapy (yes vs no)	-0.040 (0.217)	0.854	-0.025 (0.141)	0.859	-0.041 (0.142)	0.775	0.000 (0.143)	1.000	-0.031 (0.141)	0.829
PC1 (per SD)	0.057 (0.057)	0.320	0.058 (0.036)	0.107	0.058 (0.036)	0.111	0.022 (0.042)	0.596	0.059 (0.036)	0.104
PC2 (per SD)	-0.032 (0.059)	0.594	-0.005 (0.035)	0.887	-0.004 (0.035)	0.906	0.000 (0.038)	1.000	-0.005 (0.035)	0.897
PC3 (per SD)	0.035 (0.050)	0.484	-0.016 (0.034)	0.645	-0.016 (0.034)	0.639	0.000 (0.040)	1.000	-0.016 (0.034)	0.651
SNPs (ordered by SMC-DBL p-value)										
rs11022690 (C)	-0.129 (0.070)	0.066	-0.113 (0.049)	0.022	-0.112 (0.049)	0.024	-0.054 (0.052)	0.300	-0.114 (0.049)	0.022
rs72743477 (G)	-0.196 (0.090)	0.029	-0.126 (0.061)	0.038	-0.127 (0.061)	0.037	-0.021 (0.062)	0.732	-0.126 (0.061)	0.038
rs72490631 (C)	0.084 (0.084)	0.316	-0.107 (0.060)	0.074	-0.108 (0.060)	0.072	-0.017 (0.063)	0.789	-0.108 (0.060)	0.071
rs196025 (A)	-0.120 (0.074)	0.104	-0.086 (0.052)	0.099	-0.089 (0.053)	0.091	-0.028 (0.055)	0.602	-0.090 (0.052)	0.083
rs4233430 (T)	-0.082 (0.099)	0.409	-0.108 (0.067)	0.105	-0.109 (0.067)	0.105	-0.002 (0.068)	0.982	-0.112 (0.067)	0.095
rs17032590 (G)	-0.096 (0.090)	0.290	-0.102 (0.063)	0.107	-0.099 (0.063)	0.117	-0.009 (0.066)	0.890	-0.105 (0.064)	0.100
rs77972916 (A)	-0.290 (0.163)	0.075	-0.180 (0.117)	0.126	-0.182 (0.117)	0.121	0.000 (0.110)	1.000	-0.183 (0.117)	0.118
rs9660890 (C)	-0.086 (0.089)	0.338	-0.089 (0.061)	0.144	-0.089 (0.061)	0.146	-0.003 (0.063)	0.958	-0.090 (0.061)	0.141
rs7196853 (C)	0.088 (0.134)	0.512	-0.124 (0.089)	0.162	-0.127 (0.088)	0.147	-0.000 (0.088)	1.000	-0.126 (0.088)	0.152
rs72811372 (A)	-0.103 (0.145)	0.478	-0.129 (0.096)	0.177	-0.137 (0.096)	0.153	0.000 (0.092)	1.000	-0.132 (0.095)	0.164
rs11118683 (T)	0.067 (0.074)	0.363	0.067 (0.051)	0.187	0.067 (0.051)	0.186	0.002 (0.055)	0.971	0.067 (0.051)	0.185
rs17387279 (G)	0.137 (0.090)	0.128	0.077 (0.063)	0.218	0.079 (0.063)	0.209	0.000 (0.069)	1.000	0.077 (0.063)	0.222
rs28517513 (T)	0.050 (0.090)	0.582	0.067 (0.061)	0.271	0.064 (0.061)	0.295	0.000 (0.065)	1.000	0.068 (0.061)	0.260
rs1956028 (C)	0.177 (0.112)	0.116	0.081 (0.075)	0.284	0.082 (0.075)	0.278	0.000 (0.079)	1.000	0.080 (0.075)	0.290
rs11227223 (T)	0.017 (0.163)	0.917	0.119 (0.112)	0.287	0.121 (0.112)	0.279	0.000 (0.115)	1.000	0.119 (0.112)	0.289
rs4948502 (C)	0.040 (0.073)	0.586	-0.049 (0.052)	0.342	-0.048 (0.052)	0.348	0.000 (0.053)	1.000	-0.049 (0.052)	0.342

Continued on next page

Table 1 continued

Variable	CC		SMC-DBL		Std SMC-FCS		IRO		Mean Imp	
	$\hat{\beta}$ (SE)	p	$\hat{\beta}$ (SE)	p	$\hat{\beta}$ (SE)	p	$\hat{\beta}$ (SE)	p	$\hat{\beta}$ (SE)	p
rs10878300 (T)	0.078 (0.091)	0.396	0.060 (0.066)	0.362	0.059 (0.066)	0.374	0.000 (0.070)	1.000	0.061 (0.066)	0.354
rs4233284 (G)	0.040 (0.078)	0.610	0.043 (0.052)	0.409	0.041 (0.052)	0.433	0.000 (0.056)	1.000	0.044 (0.052)	0.401
rs10173269 (G)	-0.060 (0.077)	0.431	-0.042 (0.051)	0.411	-0.046 (0.051)	0.368	-0.000 (0.053)	0.999	-0.044 (0.052)	0.393
rs12466981 (T)	0.016 (0.081)	0.847	-0.042 (0.055)	0.439	-0.044 (0.055)	0.426	0.000 (0.059)	1.000	-0.043 (0.055)	0.428
rs10184235 (G)	0.158 (0.089)	0.075	0.044 (0.057)	0.443	0.046 (0.057)	0.421	0.000 (0.061)	1.000	0.044 (0.058)	0.444
rs1528624 (G)	-0.024 (0.071)	0.739	0.035 (0.049)	0.467	0.039 (0.049)	0.425	0.000 (0.052)	1.000	0.036 (0.049)	0.461
rs4800410 (C)	0.076 (0.078)	0.326	0.034 (0.053)	0.527	0.038 (0.055)	0.482	0.000 (0.056)	1.000	0.035 (0.053)	0.511
rs6006399 (G)	-0.104 (0.137)	0.450	0.051 (0.085)	0.547	0.050 (0.085)	0.552	0.000 (0.088)	1.000	0.052 (0.085)	0.543
rs7443323 (A)	-0.007 (0.080)	0.935	0.034 (0.056)	0.549	0.031 (0.056)	0.580	0.000 (0.061)	1.000	0.034 (0.056)	0.551
rs4886509 (C)	-0.015 (0.080)	0.849	0.031 (0.054)	0.562	0.034 (0.054)	0.531	0.000 (0.056)	1.000	0.035 (0.054)	0.521
rs17821105 (T)	-0.035 (0.101)	0.732	-0.039 (0.069)	0.575	-0.040 (0.069)	0.560	0.000 (0.071)	1.000	-0.039 (0.070)	0.577
rs11745375 (T)	0.125 (0.073)	0.088	0.023 (0.050)	0.641	0.024 (0.050)	0.627	0.000 (0.054)	1.000	0.023 (0.050)	0.645
rs4889526 (A)	-0.144 (0.080)	0.073	-0.024 (0.053)	0.643	-0.026 (0.053)	0.621	0.000 (0.056)	1.000	-0.025 (0.053)	0.634
rs9819463 (C)	-0.093 (0.091)	0.306	0.028 (0.062)	0.654	0.029 (0.062)	0.634	0.000 (0.068)	1.000	0.029 (0.062)	0.643
rs10987386 (T)	0.076 (0.104)	0.464	0.030 (0.067)	0.655	0.030 (0.068)	0.657	0.000 (0.071)	1.000	0.030 (0.067)	0.655
rs1548029 (C)	0.024 (0.074)	0.748	0.023 (0.051)	0.657	0.023 (0.051)	0.654	0.000 (0.055)	1.000	0.023 (0.051)	0.653
rs12313454 (G)	-0.007 (0.109)	0.952	-0.033 (0.076)	0.670	-0.033 (0.077)	0.664	0.000 (0.080)	1.000	-0.033 (0.077)	0.670
rs12571363 (A)	-0.154 (0.122)	0.207	-0.035 (0.083)	0.671	-0.035 (0.083)	0.673	0.000 (0.088)	1.000	-0.036 (0.083)	0.666
rs58453446 (C)	0.013 (0.078)	0.869	0.021 (0.053)	0.696	0.020 (0.053)	0.704	0.000 (0.057)	1.000	0.021 (0.053)	0.696
rs2794359 (A)	0.168 (0.146)	0.252	0.039 (0.100)	0.698	0.046 (0.101)	0.651	0.000 (0.104)	1.000	0.045 (0.100)	0.652
rs55993676 (T)	-0.041 (0.083)	0.624	0.019 (0.056)	0.738	0.017 (0.056)	0.759	0.000 (0.062)	1.000	0.018 (0.056)	0.749
rs1561073 (T)	0.040 (0.083)	0.634	0.017 (0.056)	0.753	0.019 (0.056)	0.735	0.000 (0.061)	1.000	0.020 (0.056)	0.718
rs659398 (T)	0.126 (0.084)	0.134	0.018 (0.058)	0.756	0.020 (0.058)	0.724	0.000 (0.060)	1.000	0.018 (0.058)	0.756
rs7927422 (C)	-0.053 (0.071)	0.450	-0.015 (0.050)	0.767	-0.016 (0.050)	0.757	0.000 (0.053)	1.000	-0.014 (0.050)	0.781
rs4444235 (T)	-0.010 (0.073)	0.896	0.014 (0.049)	0.771	0.014 (0.049)	0.769	0.000 (0.055)	1.000	0.014 (0.049)	0.771
rs28719767 (C)	-0.109 (0.080)	0.172	-0.015 (0.055)	0.782	-0.014 (0.055)	0.795	0.000 (0.059)	1.000	-0.017 (0.056)	0.759
rs9393688 (T)	-0.133 (0.083)	0.111	-0.015 (0.058)	0.795	-0.016 (0.058)	0.778	0.000 (0.060)	1.000	-0.013 (0.059)	0.827
rs1425794 (T)	0.007 (0.072)	0.920	0.012 (0.049)	0.810	0.011 (0.049)	0.819	0.000 (0.053)	1.000	0.012 (0.049)	0.810
rs2125126 (A)	0.108 (0.102)	0.291	-0.012 (0.069)	0.860	-0.010 (0.069)	0.881	0.000 (0.073)	1.000	-0.011 (0.070)	0.869

Continued on next page

Table 1 continued

Variable	CC		SMC-DBL		Std SMC-FCS		IRO		Mean Imp	
	$\hat{\beta}$ (SE)	p	$\hat{\beta}$ (SE)	p	$\hat{\beta}$ (SE)	p	$\hat{\beta}$ (SE)	p	$\hat{\beta}$ (SE)	p
rs11887136 (A)	0.197 (0.113)	0.081	0.014 (0.078)	0.860	0.017 (0.078)	0.826	0.000 (0.081)	1.000	0.014 (0.078)	0.860
rs207672 (T)	0.064 (0.076)	0.399	0.007 (0.053)	0.901	0.004 (0.053)	0.936	0.000 (0.055)	1.000	0.007 (0.053)	0.896
rs113638840 (G)	-0.003 (0.086)	0.969	-0.007 (0.061)	0.911	-0.005 (0.060)	0.929	0.000 (0.064)	1.000	-0.007 (0.060)	0.904
rs750739 (G)	0.076 (0.105)	0.470	0.008 (0.076)	0.917	0.006 (0.076)	0.933	0.000 (0.077)	1.000	0.005 (0.076)	0.947
rs4636990 (G)	-0.063 (0.070)	0.372	0.004 (0.048)	0.925	0.005 (0.048)	0.913	0.000 (0.050)	1.000	0.005 (0.048)	0.921
rs1179500 (C)	-0.049 (0.083)	0.554	-0.000 (0.057)	0.996	-0.000 (0.057)	0.993	0.000 (0.059)	1.000	0.001 (0.057)	0.984

6 Discussion

This paper extends Cox regression inference to settings with a diverging number of covariates subject to missingness, addressing an increasingly common challenge in modern biomedical and genomic survival studies. The central methodological strategy is modular; we combine Cox-compatible semiparametric multiple imputation through substantive-model-compatible fully conditional specification (SMC-FCS) with rejection sampling, followed by debiased lasso inference within each completed dataset and final combination through Rubin's rules. This separation of imputation and inference allows the method to preserve the semiparametric structure of the Cox model while maintaining valid statistical inference in moderately high-dimensional settings.

A key advantage of the proposed approach is that it avoids reliance on restrictive parametric models for the conditional distribution of missing covariates. Instead, compatibility with the Cox proportional hazards model is enforced through acceptance probabilities derived from the Cox likelihood contribution. This substantially reduces the risk of incompatibility-induced bias that may arise when standard fully conditional imputation models are used with semiparametric survival outcomes. The resulting procedure remains computationally feasible while accommodating censoring, complex covariate dependence structures, and moderate dimensionality.

The theoretical justification integrates several complementary components. First, the procedure builds on recent asymptotic theory for debiased lasso estimators in Cox regression under diverging dimensionality. Second, it leverages large-sample properties of compatible multiple imputation procedures for semiparametric survival models. Third, Rubin's variance decomposition provides a principled mechanism for propagating uncertainty arising from both imputation and penalized estimation. Together, these ingredients support asymptotically valid estimation and inference under regularity conditions when the number of covariates increases with sample size.

Several limitations and future research directions remain. First, the current method assumes that covariates are missing at random (MAR). Although MAR is standard in the multiple imputation literature, it may be unrealistic in many biomedical studies where missingness depends on unobserved patient characteristics, latent disease severity, or unmeasured clinical decisions. Future work should therefore investigate extensions under missing-not-at-random (MNAR) mechanisms. Possible directions include selection models, pattern-mixture formulations, shared-parameter models, or sensitivity-analysis tailored to high-dimensional survival settings. Developing semiparametric compatibility conditions under MNAR mechanisms would be particularly important for ensuring robust inference in observational cancer studies and electronic health record data.

Second, the current theory primarily focuses on the diverging-dimensional regime where the number of covariates grows with sample size but remains smaller than the effective sample size. Extending the method to high or ultra-high-dimensional settings with $p \gg n$ represents an important next step. Such extensions would require additional methodological developments at both the imputation and inference stages. For example, imputing high-dimensional covariates may require sparse graphical models, low-rank latent factor structures, or regularized conditional generators to stabilize estimation. On the inference side, stronger sparsity assumptions,

improved precision matrix estimation, or sample-splitting and cross-fitting strategies may be needed to maintain valid debiasing properties under extreme dimensionality. The interaction between penalized survival estimation and multiply imputed high-dimensional data also raises new theoretical questions regarding uniform convergence, error propagation across imputations, and post-selection validity.

Third, the theory assumes uniformly bounded covariates and linear predictors, which may be restrictive for continuous biomedical covariates. These assumptions facilitate concentration of the Cox partial-likelihood score and information matrix and control the remainder terms in the debiasing expansion. Also they provide uniform control of the exponential risk scores and Cox likelihood contributions needed to establish global minorization and geometric ergodicity of the SMC-FCS chains. Extending the theory to unbounded covariates would require stronger tail conditions and refined concentration arguments, particularly because exponential risk scores may exhibit heavy-tailed behavior even under sub-Gaussian covariates. Moreover, unbounded covariate support may invalidate the uniform minorization condition and destabilize the conditional imputation distributions, necessitating alternative stability arguments, such as drift-and-minorization or localization techniques. These challenges complicate the establishment of uniform convergence rates and asymptotic normality as the covariate dimension diverges. Relaxing the boundedness assumptions while preserving both valid high-dimensional inference and convergence guarantees for the imputation procedure remains an important direction for future research.

Additional extensions include time-dependent covariates, competing risks, recurrent events, and clustered survival outcomes. These settings would require modification of both the substantive-model contribution used in the imputation step and the corresponding inferential theory. Incorporating more flexible machine-learning-based conditional imputation models while retaining substantive-model compatibility is another promising direction, particularly when complex nonlinear dependence among covariates cannot be adequately represented by the current working models. More broadly, the proposed framework provides a foundation for combining principled missing-data handling with modern high-dimensional survival inference, with potential applications in genomics, radiomics, multimodal cancer studies, and large-scale electronic health record analyses.

Acknowledgements

We thank the Editor, Associate Editor, and two referees for their insightful comments, which have substantially improved the presentation of the paper.

References

Jonathan W. Bartlett, Shaun R. Seaman, Ian R. White, and James R. Carpenter. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive mode. *Statistical Methods in Medical Research*, 24(4):462–487, 2015. doi: 10.1177/0962280214521348.

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Jianqing Fan and Runze Li. Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, 30(1):74–99, 2002.

Ethan X. Fang, Yang Ning, and Han Liu. Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1415–1437, 2017. doi: 10.1111/rssb.12224.

Jian Huang, Tingni Sun, Zhiliang Ying, Yi Yu, and Cun-Hui Zhang. Oracle inequalities for the lasso in the Cox model. *Annals of Statistics*, 41(3):1142–1165, 2013.

Rachael A. Hughes, Ian R. White, Shaun R. Seaman, James R. Carpenter, Kate Tilling, and Jonathan A. C. Sterne. Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14:28, 2014. doi: 10.1186/1471-2288-14-28.

Linda Kachuri, Mattias Johansson, Sara R. Rashkin, Rebecca E. Graff, Yohan Bossé, Venkata Manem, Neil E. Caporaso, Maria Teresa Landi, David C. Christiani, Paolo Vineis, et al. Immune-mediated genetic pathways resulting in pulmonary function impairment increase lung cancer susceptibility. *Nature Communications*, 11:27, 2020. doi: 10.1038/s41467-019-13855-2.

Ruth H. Keogh and Tim P. Morris. Multiple imputation in Cox regression when there are time-varying effects of covariates. *Statistics in Medicine*, 37(25):3661–3678, 2018. doi: 10.1002/sim.7842.

Ruth H. Keogh, Shaun R. Seaman, Jonathan W. Bartlett, and Angela M. Wood. Multiple imputation of missing data in nested case-control and case-cohort studies. *Biometrics*, 74(4):1438–1449, 2018. doi: 10.1111/biom.12910.

Shengchun Kong, Zhuqing Yu, Xianyang Zhang, and Guang Cheng. High-dimensional robust inference for Cox regression models using desparsified lasso. *Scandinavian Journal of Statistics*, 48(3):1068–1095, 2021. doi: 10.1111/sjos.12543.

Faming Liang, Bochao Jia, Jingnan Xue, Qizhai Li, and Ye Luo. An imputation-regularized optimization algorithm for high dimensional missing data problems and beyond. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):899–926, 2018. doi: 10.1111/rssb.12279.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Hoboken, NJ, 3 edition, 2019.

Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558, 1994.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

Robert Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

Stef Van Buuren, Jaap P. L. Brand, Catharina G. M. Groothuis-Oudshoorn, and Donald B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, 2006. doi: 10.1080/10629360600810434.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3): 1166–1202, 2014. doi: 10.1214/14-AOS1221.

Naisyin Wang and James M. Robins. Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85(4):935–948, 1998.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional EM algorithm: statistical optimization and asymptotic normality. In *Advances in Neural Information Processing Systems*, volume 28, pages 2521–2529, 2015.

Ian R. White and Patrick Royston. Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28(15):1982–1998, 2009.

Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011.

Lu Xia, Bin Nan, and Yi Li. Statistical inference for Cox proportional hazards models with a diverging number of covariates. *Scandinavian Journal of Statistics*, 50(2):550–571, 2023. doi: 10.1111/sjos.12595.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014. doi: 10.1111/rssb.12026.

7 Appendix: Proofs, Demographics of BLCS and Additional Simulation Results

This section presents technical proofs of Theorems 1–4, descriptive analysis of BLCS and additional numerical evidence supporting the main results of the paper.

7.1 Proofs

Proof of Theorem 1. Fix β satisfying $\|\beta - \beta^0\|_1 \leq r$. By Assumptions 1 and 2,

$$|X_i^\top \beta| \leq |X_i^\top \beta^0| + \|X_i\|_\infty \|\beta - \beta^0\|_1 \leq K_1 + Kr \equiv B \tag{S.1}$$

uniformly in i. Thus, $e^{-B} \leq \exp(X_i^\top \beta) \leq e^B$ for all i.

Let $\widehat{\Lambda}_0$ denote the Breslow estimator based on the current completed data,

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{dN(u)}{\sum_{i=1}^n I(Y_i \geq u) \exp(X_i^\top \beta)},$$

where $N(u) = \sum_{i=1}^n I(Y_i \leq u, \Delta_i = 1)$. By Assumption 3, $\Pr(Y \geq \tau) = \pi_0 > 0$. Therefore, by the law of large numbers, with probability tending to one,

$$n^{-1} \sum_{i=1}^n I(Y_i \geq \tau) \geq \pi_0/2. \tag{S.2}$$

On the event in (S.2), for every $u \leq \tau$, monotonicity of the risk sets and (S.1) imply

$$\sum_{i=1}^n I(Y_i \geq u) \exp(X_i^\top \beta) \geq \sum_{i=1}^n I(Y_i \geq \tau) \exp(X_i^\top \beta) \geq n e^{-B} \pi_0/2. \tag{S.3}$$

Combining the Breslow representation and (S.3), and using $N(\tau) \leq n$, gives

$$0 \leq \widehat{\Lambda}_0(\tau) \leq \frac{N(\tau)}{n e^{-B} \pi_0/2} \leq \frac{2e^B}{\pi_0} \equiv C_\Lambda. \tag{S.4}$$

Similarly, at any observed failure time $u \leq \tau$,

$$0 \leq \Delta\widehat{\Lambda}_0(u) = \frac{dN(u)}{\sum_{i=1}^n I(Y_i \geq u) \exp(X_i^\top \beta)} \leq \frac{\sum_{i=1}^n I(Y_i \geq u)}{e^{-B} \sum_{i=1}^n I(Y_i \geq u)} = e^B \equiv C_\Delta.$$

Thus the boundedness of the Breslow cumulative hazard and its jumps follows from the stated assumptions with probability tending to one.

Now again condition on the high-probability event on which (S.2) holds. The inner SMC-FCS chain updates only the missing covariate entries, so its state space is $\mathcal{X} = \prod_{(i,j):R_{ij}=0} \mathcal{X}_{ij}$, with total dimension d.

For a missing coordinate (i, j), write x for a candidate value of X_{ij} and $X_i(x)$ for the full covariate vector obtained by replacing X_{ij} by x. In the Metropolis–Hastings ratio, the covariate-dependent part of the Cox full-likelihood contribution is

$$g_i(x; \beta, \widehat{\Lambda}_0) = \exp\left\{\Delta_i X_i(x)^\top \beta - \widehat{\Lambda}_0(Y_i) \exp(X_i(x)^\top \beta)\right\}.$$

The jump factor $\{\Delta\widehat{\Lambda}_0(Y_i)\}^{\Delta_i}$ is omitted because it does not depend on x and cancels in the Metropolis–Hastings ratio. By (S.1) and (S.4), there exist constants $0 < c_f \leq C_f < \infty$ such that

$$c_f \leq g_i(x; \beta, \widehat{\Lambda}_0) \leq C_f \tag{S.5}$$

uniformly over all admissible x, i, and j.

By Assumption 6, the proposal density or mass function q_{ij} satisfies $c_\theta \leq q_{ij}(x) \leq C_\theta$ uniformly over its support. The one-coordinate target distribution is proportional to

$$q_{ij}(x) g_i(x; \beta, \widehat{\Lambda}_0).$$

Using (S.5) and the bounds on q_{ij}, the proposal and target are uniformly equivalent, and hence there exists

$$\varepsilon_a = \frac{c_\theta c_f}{C_\theta C_f} \in (0,1)$$

such that, for every current value x_{curr},

$$K_{ij}(x_{\text{curr}}, \cdot) \geq \varepsilon_a \, \pi_{ij}(\cdot \mid X_{-(ij)}; \beta, \widehat{\Lambda}_0), \tag{S.6}$$

where $\pi_{ij}(\cdot \mid X_{-(ij)}; \beta, \widehat{\Lambda}_0)$ denotes the normalized one-coordinate conditional distribution induced by the above target density. The same argument applies to discrete coordinates, with sums replacing integrals.

A full SMC-FCS sweep is the composition of the d one-coordinate kernels. Repeatedly applying (S.6) over the d missing coordinates yields the Doeblin minorisation

$$P_{\beta,\widehat{\Lambda}_0}(x, \cdot) \geq \varepsilon_a^d \mu_{\beta,\widehat{\Lambda}_0}(\cdot)$$

for some probability measure $\mu_{\beta,\widehat{\Lambda}_0}$ on $\mathcal{X}$. Therefore $P_{\beta,\widehat{\Lambda}_0}$ is uniformly ergodic and admits a unique stationary distribution $\nu_{\beta,\widehat{\Lambda}_0}$. Moreover, by the Doeblin bound,

$$\left\| P^S_{\beta,\widehat{\Lambda}_0}(x_0, \cdot) - \nu_{\beta,\widehat{\Lambda}_0} \right\|_{\text{TV}} \leq (1 - \varepsilon_a^d)^S. \tag{S.7}$$

Thus the theorem holds with $C_0 = 1$ and $\rho_d = 1 - \varepsilon_a^d$; allowing a generic constant $C_0 < \infty$ gives the stated form.

Finally, $C_0(1 - \varepsilon_a^d)^S \leq \delta$ is implied by $S \geq \log(C_0/\delta)/\varepsilon_a^d$, since $\log(1-u) \leq -u$ for $u \in (0,1)$. This completes the proof. □

Proof of Theorem 2. Let $\theta = (\beta, \Lambda_0)$ and let $\theta^* = (\beta^*, \Lambda_0^*)$ denote the fixed point of the population IRO map. Define

$$\|\theta - \theta^*\|_{\mathcal{H}} = \|\beta - \beta^*\|_1 + \|\Lambda_0 - \Lambda_0^*\|_\infty.$$

Let $\theta^{(\ell)} = (\widetilde{\beta}^{(\ell)}, \widehat{\Lambda}_0^{(\ell)})$ denote the outer IRO iterate at iteration ℓ. The sample outer update can be written as

$$\theta^{(\ell+1)} = \widehat{\mathcal{M}}_n(\theta^{(\ell)}),$$

where $\widehat{\mathcal{M}}_n$ is the empirical map induced by: first running S_{in} inner SMC-FCS sweeps conditional on $\theta^{(\ell)}$, then refitting the Cox lasso, and finally updating the Breslow baseline cumulative hazard. Let $\mathcal{M}$ denote the corresponding population map based on the stationary SMC-FCS law.

By Assumption 12, there exist a neighbourhood $\mathcal{N}(\theta^*)$, a constant $\lambda_* \in (0,1)$, and a deterministic sequence $a_n = o(1)$ such that, for every $\theta \in \mathcal{N}(\theta^*)$,

$$\|\mathcal{M}(\theta) - \theta^*\|_{\mathcal{H}} \leq \lambda_* \|\theta - \theta^*\|_{\mathcal{H}} + a_n. \tag{S.8}$$

We next show that the empirical map $\widehat{\mathcal{M}}_n$ uniformly approximates $\mathcal{M}$ in a neighbourhood of θ^*. By Theorem 1 and Assumption 10, the finite-sweep error from the inner SMC-FCS chain is $o(1)$ uniformly over $\mathcal{N}(\theta^*)$, so the completed-data objective generated by the finite inner chain

differs from its stationary counterpart by $o_p(1)$. We establish this separately for the regression and baseline components.

For the regression component, write the completed-data empirical objective as

$$Q_n(b;\theta) = \ell_n\{b;\widetilde{X}(\theta)\} + \lambda_n\|b\|_1,$$

where $\widetilde{X}(\theta)$ denotes a completed dataset generated from the inner SMC-FCS law conditional on $\theta = (\beta, \Lambda_0)$. Let ν_θ denote the stationary SMC-FCS distribution of the missing covariates conditional on the observed data and substantive-model parameter $\theta = (\beta, \Lambda_0)$. Let $\ell_n^{\mathrm{comp}}(b;\widetilde{X})$ denote the completed-data negative Cox partial log-likelihood evaluated at coefficient vector b for completed dataset $\widetilde{X}$. Define the population objective

$$Q(b;\theta) = \mathbb{E}_{\nu_\theta}\big[\ell_n^{\mathrm{comp}}(b;\widetilde{X})\big] + \lambda_n\|b\|_1,$$

where the expectation is taken with respect to the stationary imputation law ν_θ. Then

$$\widehat{\mathcal{M}}_{n,\beta}(\theta) = \arg\min_b Q_n(b;\theta), \qquad \mathcal{M}_\beta(\theta) = \arg\min_b Q(b;\theta).$$

We first control the difference between Q_n and Q. Under Assumptions 1–7, the completed covariates are uniformly bounded, the risk-set denominators are bounded away from zero on $[0,\tau]$, and the Cox partial log-likelihood is locally Lipschitz in b. In particular, for $b, b' \in \mathcal{N}_\beta(\beta^*)$,

$$\big|\ell_n(b;\widetilde{X}) - \ell_n(b';\widetilde{X})\big| \le C\|b - b'\|_1$$

with probability tending to one, uniformly over completed datasets generated in a local neighbourhood of θ^*. Standard stochastic equicontinuity for bounded empirical processes therefore gives

$$\sup_{b\in\mathcal{N}_\beta(\beta^*),\,\theta\in\mathcal{N}(\theta^*)} \big|Q_n(b;\theta) - Q(b;\theta)\big| = o_p(1),$$

where the finite-sweep error is absorbed using Theorem 1 and Assumption 10.

Next, the Cox lasso objective is convex in b, and Assumption 5 gives local curvature of the population risk around $\mathcal{M}_\beta(\theta)$. Hence, for some $\kappa > 0$ and all b in the local neighbourhood,

$$Q(b;\theta) - Q\{\mathcal{M}_\beta(\theta);\theta\} \ge \kappa\|b - \mathcal{M}_\beta(\theta)\|_2^2 - o(1),$$

uniformly over $\theta \in \mathcal{N}(\theta^*)$. The uniform convergence of Q_n to Q, together with the argmin stability theorem for convex functions, implies

$$\sup_{\theta\in\mathcal{N}(\theta^*)} \|\widehat{\mathcal{M}}_{n,\beta}(\theta) - \mathcal{M}_\beta(\theta)\|_2 = o_p(1).$$

Since the iterates are restricted to the local sparse neighbourhood by the bounded-moment condition $\sup_{n,\ell}\mathbb{E}\|\widetilde{\beta}^{(\ell)}\|_1 < \infty$ and the local sparsity assumptions, the same conclusion holds in ℓ_1 norm:

$$\sup_{\theta\in\mathcal{N}(\theta^*)} \|\widehat{\mathcal{M}}_{n,\beta}(\theta) - \mathcal{M}_\beta(\theta)\|_1 = o_p(1). \tag{S.9}$$

For the baseline component, recall that for a completed dataset and coefficient vector b, the Breslow estimator is

$$\widehat{\Lambda}_b(t) = \sum_{r:Y_r \le t, \Delta_r = 1} \left[S_n^{(0)}(Y_r; b) \right]^{-1},$$

where

$$S_n^{(0)}(u; b) = \sum_{i:Y_i \ge u} \exp(X_i^\top b)$$

denotes the empirical risk-set denominator.

Let b, b' belong to a neighbourhood $\mathcal{N}_\beta(\beta^*)$. By Assumptions 1 and 2, there exists $B < \infty$ such that

$$|X_i^\top b| \le B, \qquad |X_i^\top b'| \le B$$

uniformly in i with probability tending to one. Hence

$$e^{-B} \le \exp(X_i^\top b), \exp(X_i^\top b') \le e^B.$$

By Assumption 3, the risk sets remain nondegenerate on $[0, \tau]$, and therefore

$$n^{-1} S_n^{(0)}(u; b) \ge c_0$$

uniformly over $u \le \tau$ and $b \in \mathcal{N}_\beta(\beta^*)$ with probability tending to one, for some constant $c_0 > 0$. Consequently,

$$S_n^{(0)}(u; b) \ge n c_0 \tag{S.10}$$

uniformly over $u \le \tau$ and $b \in \mathcal{N}_\beta(\beta^*)$.

Next, using the mean-value theorem,

$$\left| \exp(X_i^\top b) - \exp(X_i^\top b') \right| = \exp(X_i^\top \widetilde{b}) \left| X_i^\top (b - b') \right|$$

for some $\widetilde{b}$ on the line segment joining b and b'. Since $\|X_i\|_\infty \le K$ and $\exp(X_i^\top \widetilde{b}) \le e^B$,

$$\left| \exp(X_i^\top b) - \exp(X_i^\top b') \right| \le K e^B \|b - b'\|_1.$$

Summing over the risk set gives

$$\left| S_n^{(0)}(u; b) - S_n^{(0)}(u; b') \right| \le n K e^B \|b - b'\|_1.$$

Now consider the difference between the two Breslow estimators:

$$\begin{aligned} \left| \widehat{\Lambda}_b(t) - \widehat{\Lambda}_{b'}(t) \right| &= \left| \sum_{r:Y_r \le t, \Delta_r = 1} \left[\frac{1}{S_n^{(0)}(Y_r; b)} - \frac{1}{S_n^{(0)}(Y_r; b')} \right] \right| \\ &= \left| \sum_{r:Y_r \le t, \Delta_r = 1} \frac{S_n^{(0)}(Y_r; b') - S_n^{(0)}(Y_r; b)}{S_n^{(0)}(Y_r; b)\, S_n^{(0)}(Y_r; b')} \right|. \end{aligned}$$

Using (S.10),

$$S_n^{(0)}(Y_r; b)\, S_n^{(0)}(Y_r; b') \ge n^2 c_0^2,$$

while the numerator is bounded by

$$nKe^B\|b-b'\|_1.$$

Since the number of failure times is at most n,

$$|\widehat{\Lambda}_b(t)-\widehat{\Lambda}_{b'}(t)| \le \sum_{r:Y_r\le t,\Delta_r=1} \frac{nKe^B\|b-b'\|_1}{n^2c_0^2} \le \frac{Ke^B}{c_0^2}\|b-b'\|_1.$$

Taking the supremum over $t \le \tau$ yields

$$\|\widehat{\Lambda}_b-\widehat{\Lambda}_{b'}\|_\infty \le C\|b-b'\|_1 \tag{S.11}$$

with probability tending to one, where

$$C=\frac{Ke^B}{c_0^2}.$$

Thus the Breslow update is locally Lipschitz in the regression coefficient over the neighbourhood $\mathcal{N}_\beta(\beta^*)$. Combining this Lipschitz property with the uniform convergence of the empirical risk-set processes gives

$$\sup_{\theta\in\mathcal{N}(\theta^*)} \|\widehat{\mathcal{M}}_{n,\Lambda}(\theta)-\mathcal{M}_\Lambda(\theta)\|_\infty = o_p(1). \tag{S.12}$$

Combining (S.9) and (S.12), we obtain the uniform approximation

$$\sup_{\theta\in\mathcal{N}(\theta^*)} \|\widehat{\mathcal{M}}_n(\theta)-\mathcal{M}(\theta)\|_{\mathcal{H}} = o_p(1). \tag{S.13}$$

Now consider one outer update. On the high-probability event that $\theta^{(\ell)}\in\mathcal{N}(\theta^*)$, we have

$$\begin{aligned}\|\theta^{(\ell+1)}-\theta^*\|_{\mathcal{H}} &= \|\widehat{\mathcal{M}}_n(\theta^{(\ell)})-\theta^*\|_{\mathcal{H}}\\ &\le \|\widehat{\mathcal{M}}_n(\theta^{(\ell)})-\mathcal{M}(\theta^{(\ell)})\|_{\mathcal{H}} + \|\mathcal{M}(\theta^{(\ell)})-\theta^*\|_{\mathcal{H}}.\end{aligned}$$

Using (S.8) and (S.13),

$$\|\theta^{(\ell+1)}-\theta^*\|_{\mathcal{H}} \le \lambda_*\|\theta^{(\ell)}-\theta^*\|_{\mathcal{H}} + a_n + o_p(1). \tag{S.14}$$

Iterating (S.14) gives

$$\|\theta^{(\ell)}-\theta^*\|_{\mathcal{H}} \le \lambda_*^\ell\|\theta^{(0)}-\theta^*\|_{\mathcal{H}} + \sum_{q=0}^{\ell-1}\lambda_*^q\{a_n+o_p(1)\}.$$

Since $\sum_{q=0}^{\ell-1}\lambda_*^q \le (1-\lambda_*)^{-1}$, it follows that

$$\|\theta^{(\ell)}-\theta^*\|_{\mathcal{H}} \le \lambda_*^\ell\|\theta^{(0)}-\theta^*\|_{\mathcal{H}} + \frac{a_n+o_p(1)}{1-\lambda_*}. \tag{S.15}$$

By Assumption 11,

$$\ell \ge c_\ell\frac{\log n}{\log(1/\lambda_*)}$$

with $c_\ell > 1$. Hence $\lambda_*^\ell = O(n^{-c_\ell}) = o(1)$. The bounded-moment condition on $\widetilde{\beta}^{(\ell)}$, together with the boundedness of the Breslow estimator derived from Assumptions 1, 2, and 3, implies

$$\|\theta^{(0)} - \theta^*\|_{\mathcal{H}} = O_p(1).$$

Therefore,

$$\lambda_*^\ell \|\theta^{(0)} - \theta^*\|_{\mathcal{H}} = o_p(1).$$

Since $a_n = o(1)$, the second term in (S.15) is also $o_p(1)$. Thus

$$\|\theta^{(\ell)} - \theta^*\|_{\mathcal{H}} = o_p(1).$$

Finally, substituting

$$\theta^{(\ell)} = (\widetilde{\beta}^{(\ell)}, \widehat{\Lambda}_0^{(\ell)}) \qquad \text{and} \qquad \theta^* = (\beta^*, \Lambda_0^*)$$

into the definition of $\|\cdot\|_{\mathcal{H}}$ gives

$$\|\widetilde{\beta}^{(\ell)} - \beta^*\|_1 + \|\widehat{\Lambda}_0^{(\ell)} - \Lambda_0^*\|_\infty = o_p(1).$$

This proves the theorem. □

Proof of Theorem 3. At stationarity, $\widetilde{\beta}^{(\ell)}$ is computed from a completed dataset drawn from the stationary law associated with (β^*, Λ_0^*). For notational ease, let $L_n(b) = \ell_n^{\text{comp}}(b)$ denote the completed-data empirical negative partial log-likelihood, and write

$$\widetilde{\beta}^{(\ell)} = \arg\min_{b \in \mathbb{R}^p} \{L_n(b) + \lambda_n \|b\|_1\}.$$

Let $\Delta = \widetilde{\beta}^{(\ell)} - \beta^*$ and $S = \text{supp}(\beta^*)$, with $|S| \le s_0$ up to the fixed-point perturbation allowed under the assumptions.
By optimality of $\widetilde{\beta}^{(\ell)}$,

$$L_n(\beta^* + \Delta) + \lambda_n \|\beta^* + \Delta\|_1 \le L_n(\beta^*) + \lambda_n \|\beta^*\|_1. \tag{S.16}$$

Rearranging (S.16) and using the support decomposition gives

$$L_n(\beta^* + \Delta) - L_n(\beta^*) \le \lambda_n (\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1). \tag{S.17}$$

Expand the left-hand side around β^*:

$$L_n(\beta^* + \Delta) - L_n(\beta^*) = \dot{L}_n(\beta^*)^\top \Delta + R_n(\Delta), \tag{S.18}$$

where $R_n(\Delta)$ is the empirical curvature remainder. Under Assumptions 1, 5, 8, and 9, together with the stationary completed-data law supplied by Theorem 1, the Cox score satisfies

$$\|\dot{L}_n(\beta^*)\|_\infty = O_p\left(\sqrt{\log p / n}\right). \tag{S.19}$$

With $\lambda_n \asymp \sqrt{\log p/n}$ chosen sufficiently large, (S.19) implies that $\|\dot{L}_n(\beta^*)\|_\infty \leq \lambda_n/2$ with probability tending to one. Hence

$$|\dot{L}_n(\beta^*)^\top \Delta| \leq \frac{\lambda_n}{2}\|\Delta\|_1. \tag{S.20}$$

Combining (S.17), (S.18), and (S.20), with probability tending to one,

$$R_n(\Delta) \leq \frac{\lambda_n}{2}\|\Delta\|_1 + \lambda_n(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1).$$

Since $R_n(\Delta) \geq 0$ by local convexity of the Cox partial likelihood, the preceding display yields

$$0 \leq \frac{\lambda_n}{2}(\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1) + \lambda_n(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1),$$

and hence

$$\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1. \tag{S.21}$$

Thus Δ lies in the usual lasso cone.
On this cone, Assumption 5, together with the empirical restricted-eigenvalue consequence of Assumption 9, gives the restricted curvature bound

$$R_n(\Delta) \geq \kappa\|\Delta\|_2^2 \tag{S.22}$$

for some $\kappa > 0$ with probability tending to one. Using (S.17), (S.18), (S.20), and (S.21), we obtain

$$\kappa\|\Delta\|_2^2 \leq \frac{3}{2}\lambda_n\|\Delta_S\|_1 \leq \frac{3}{2}\lambda_n\sqrt{s_0}\|\Delta\|_2.$$

Therefore,

$$\|\Delta\|_2 = O_p(\sqrt{s_0}\lambda_n). \tag{S.23}$$

By the cone condition (S.21),

$$\|\Delta\|_1 \leq 4\|\Delta_S\|_1 \leq 4\sqrt{s_0}\|\Delta\|_2.$$

Combining this with (S.23) gives

$$\|\Delta\|_1 = O_p(s_0\lambda_n).$$

Since $\lambda_n \asymp \sqrt{\log p/n}$, we obtain

$$\|\widetilde{\beta}^{(\ell)} - \beta^*\|_1 = O_p\left(s_0\sqrt{\log p/n}\right).$$

If, in addition,

$$\|\beta^* - \beta^0\|_1 = O\left(s_0\sqrt{\log p/n}\right),$$

then the triangle inequality gives

$$\|\widetilde{\beta}^{(\ell)} - \beta^0\|_1 \leq \|\widetilde{\beta}^{(\ell)} - \beta^*\|_1 + \|\beta^* - \beta^0\|_1 = O_p\left(s_0\sqrt{\log p/n}\right).$$

Finally, Theorem 2 gives

$$\|\widehat{\Lambda}_0^{(\ell)} - \Lambda_0^*\|_\infty = o_p(1),$$

which proves the result. □

Proof of Theorem 4. We prove the theorem in four steps.
First consider one retained completed dataset generated from the stationary SMC-FCS law. By Theorems 1 and 2, together with Assumptions 10 and 13, the retained completed datasets are asymptotically distributed according to the stationary imputation law associated with the limiting IRO fixed point. By Theorem 3,

$$\|\widehat{\beta}_{L_1}^{(m)} - \beta^0\|_1 = O_p(s_0\lambda_n), \qquad \lambda_n \asymp \sqrt{\log p/n}, \tag{S.24}$$

up to the fixed-point perturbation already assumed to be of the same order.
For the mth completed dataset, the debiased estimator is

$$\widehat{\beta}_{\mathrm{db}}^{(m)} = \widehat{\beta}_{L_1}^{(m)} - \widehat{\Theta}^{(m)}\dot{\ell}_n^{(m)}(\widehat{\beta}_{L_1}^{(m)}).$$

Expand the Cox score around β^0:

$$\dot{\ell}_n^{(m)}(\widehat{\beta}_{L_1}^{(m)}) = \dot{\ell}_n^{(m)}(\beta^0) + \ddot{\ell}_n^{(m)}(\beta^0)(\widehat{\beta}_{L_1}^{(m)} - \beta^0) + R_n^{(m)}, \tag{S.25}$$

where $R_n^{(m)}$ is the Taylor remainder. Under Assumptions 1, 5, 8, and 9, the Hessian process is locally Lipschitz in a neighbourhood of β^0, and therefore the remainder satisfies

$$\|R_n^{(m)}\|_\infty = o_p(n^{-1/2}). \tag{S.26}$$

Substituting (S.25) into the definition of $\widehat{\beta}_{\mathrm{db}}^{(m)}$ gives

$$\widehat{\beta}_{\mathrm{db}}^{(m)} - \beta^0 = [I_p - \widehat{\Theta}^{(m)}\ddot{\ell}_n^{(m)}(\beta^0)](\widehat{\beta}_{L_1}^{(m)} - \beta^0) - \widehat{\Theta}^{(m)}\dot{\ell}_n^{(m)}(\beta^0) - \widehat{\Theta}^{(m)}R_n^{(m)}.$$

The nodewise quadratic program and the choice $\gamma_n \asymp \|\Theta_{\beta^0}\|_{1,1}s_0\lambda_n$ imply that

$$\|\widehat{\Theta}^{(m)}\ddot{\ell}_n^{(m)}(\beta^0) - I_p\|_\infty = o_p(1) \tag{S.27}$$

on fixed-dimensional contrasts. Combining (S.24), (S.26), and (S.27) yields

$$\widehat{\beta}_{\mathrm{db}}^{(m)} - \beta^0 = -\Theta_{\beta^0}\dot{\ell}_n^{(m)}(\beta^0) + r_n^{(m)},$$

where the remainder satisfies

$$\sqrt{n}\, c^\top r_n^{(m)} = o_p(1) \tag{S.28}$$

for every fixed loading vector c satisfying $\|c\|_2 = 1$ and $\|c\|_1 \le a^*$.
The Cox score admits the martingale representation

$$\dot{\ell}_n^{(m)}(\beta^0) = \frac{1}{n}\sum_{i=1}^n \psi^{(m)}(O_i) + o_p(n^{-1/2}),$$

where $\psi^{(m)}$ is the completed-data efficient score contribution. Therefore,

$$\sqrt{n}\, c^\top(\widehat{\beta}_{\mathrm{db}}^{(m)} - \beta^0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_c^{(m)}(O_i) + o_p(1), \tag{S.29}$$

where

$$\phi_c^{(m)}(O_i) = -c^\top \Theta_{\beta^0} \psi^{(m)}(O_i)$$

is the completed-data influence function for the contrast $c^\top \beta^0$.

Second, because the SMC-FCS procedure is assumed to be asymptotically proper and congenial with the substantive Cox model, the completed-data influence function admits the standard observed-data decomposition from multiple-imputation theory. Consequently, $\bar{\beta}_M = \frac{1}{M}\sum_{m=1}^{M} \widehat{\beta}_{\mathrm{db}}^{(m)}$ satisfies

$$\sqrt{n}\, c^\top(\bar{\beta}_M - \beta^0) = \frac{1}{M} \sum_{m=1}^{M} \sqrt{n}\, c^\top(\widehat{\beta}_{\mathrm{db}}^{(m)} - \beta^0) + o_p(1).$$

Substituting (S.29),

$$\sqrt{n}\, c^\top(\bar{\beta}_M - \beta^0) = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_c^{(m)}(O_i) + o_p(1).$$

Under properness and congeniality, Rubin's variance decomposition applies. The complete-data variance is $V_{\mathrm{com}} = \Theta_{\beta^0}$, while the additional variance induced by missing covariates is $V_{\mathrm{mis}} = V_{\mathrm{ML}} - V_{\mathrm{com}}$. Hence, for fixed finite M,

$$n \operatorname{Var}(c^\top(\bar{\beta}_M - \beta^0)) \to c^\top \left[V_{\mathrm{com}} + \left(1 + \frac{1}{M}\right) V_{\mathrm{mis}} \right] c. \tag{S.30}$$

Applying the classical central limit theorem to the influence-function representation therefore gives

$$\sqrt{n}\, c^\top(\bar{\beta}_M - \beta^0) \xrightarrow{d} \mathcal{N}\left(0,\, c^\top \left[V_{\mathrm{com}} + \left(1 + \frac{1}{M}\right) V_{\mathrm{mis}} \right] c\right). \tag{S.31}$$

Third, Rubin's total variance estimator is

$$\widehat{V}_{\mathrm{total}} = \widehat{V}_W + \left(1 + \frac{1}{M}\right) \widehat{V}_B,$$

where $\widehat{V}_W$ is the average within-imputation variance and $\widehat{V}_B$ is the between-imputation covariance. Properness and congeniality imply

$$n\, c^\top \widehat{V}_W c \xrightarrow{p} c^\top V_{\mathrm{com}} c \tag{S.32}$$

and

$$n\, c^\top \widehat{V}_B c \xrightarrow{p} c^\top V_{\mathrm{mis}} c. \tag{S.33}$$

Combining (S.32) and (S.33),

$$n\, c^\top \widehat{V}_{\mathrm{total}} c \xrightarrow{p} c^\top \left[V_{\mathrm{com}} + \left(1 + \frac{1}{M}\right) V_{\mathrm{mis}} \right] c. \tag{S.34}$$

Combining (S.31) and (S.34), Slutsky's theorem yields

$$\frac{\sqrt{n}\, c^\top (\bar{\beta}_M - \beta^0)}{\sqrt{n\, c^\top \widehat{V}_{\text{total}} c}} \xrightarrow{d} \mathcal{N}(0,1).$$

□

7.2 Additional Data and Simulation Results

Table S.1: Baseline characteristics of the NSCLC cohort (N = 977), stratified by vital status at the end of follow-up. Continuous variables are summarised as mean (SD); categorical variables as n (%). Group comparisons use the two-sample t-test for continuous variables and the χ^2 test (or Fisher's exact test when an expected cell count is below five) for categorical variables. Missingness is reported on a separate row for every variable. The 51 SNPs are coded on the additive 0/1/2 dosage scale, with genotype counts and missingness shown individually.

Characteristic	Overall (N=977)	Alive (n=206)	Deceased (n=771)	p-value
Age (years), mean (SD)	65.48 (10.56)	61.28 (10.83)	66.60 (10.20)	<0.001
Female sex, n (%)	502 (51.4%)	76 (36.9%)	426 (55.3%)	<0.001
Ever-smoker, n (%)	886 (90.7%)	176 (85.4%)	710 (92.1%)	0.005
Late-stage disease, n (%)	454 (46.5%)	117 (56.8%)	337 (43.7%)	0.001
Radiotherapy, n (%)	271 (27.7%)	44 (21.4%)	227 (29.4%)	0.024
Missing, n (%)	4 (0.4%)	0 (0.0%)	4 (0.5%)	
Chemotherapy, n (%)	497 (50.9%)	91 (44.2%)	406 (52.7%)	0.031
Missing, n (%)	4 (0.4%)	0 (0.0%)	4 (0.5%)	
Surgery, n (%)	564 (57.7%)	142 (68.9%)	422 (54.7%)	<0.001
Missing, n (%)	4 (0.4%)	0 (0.0%)	4 (0.5%)	
Principal component 1, mean (SD)	0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.504
Principal component 2, mean (SD)	0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.368
Principal component 3, mean (SD)	0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.633
SNP markers (additive 0/1/2 dosage, n = 51)				
rs2794359 (A), n (%)				0.778
genotype 0	801 (82.0%)	172 (83.5%)	629 (81.6%)	
genotype 1	130 (13.3%)	25 (12.1%)	105 (13.6%)	
genotype 2	2 (0.2%)	0 (0.0%)	2 (0.3%)	
Missing, n (%)	44 (4.5%)	9 (4.4%)	35 (4.5%)	
rs9660890 (C), n (%)				0.429
genotype 0	600 (61.4%)	118 (57.3%)	482 (62.5%)	
genotype 1	308 (31.5%)	72 (35.0%)	236 (30.6%)	
genotype 2	48 (4.9%)	10 (4.9%)	38 (4.9%)	
Missing, n (%)	21 (2.1%)	6 (2.9%)	15 (1.9%)	
rs4233284 (G), n (%)				0.753
genotype 0	451 (46.2%)	91 (44.2%)	360 (46.7%)	
genotype 1	410 (42.0%)	89 (43.2%)	321 (41.6%)	
genotype 2	113 (11.6%)	26 (12.6%)	87 (11.3%)	
Missing, n (%)	3 (0.3%)	0 (0.0%)	3 (0.4%)	
rs2125126 (A), n (%)				0.770
genotype 0	708 (72.5%)	149 (72.3%)	559 (72.5%)	
genotype 1	233 (23.8%)	49 (23.8%)	184 (23.9%)	

continued on next page

Table S.1 continued from previous page

Characteristic	Overall (N=977)	Alive (n=206)	Deceased (n=771)	*p*-value
genotype 2	26 (2.7%)	7 (3.4%)	19 (2.5%)	
Missing, n (%)	10 (1.0%)	1 (0.5%)	9 (1.2%)	
rs4233430 (T), n (%)				0.476
genotype 0	642 (65.7%)	142 (68.9%)	500 (64.9%)	
genotype 1	286 (29.3%)	54 (26.2%)	232 (30.1%)	
genotype 2	28 (2.9%)	7 (3.4%)	21 (2.7%)	
Missing, n (%)	21 (2.1%)	3 (1.5%)	18 (2.3%)	
rs11118683 (T), n (%)				0.960
genotype 0	338 (34.6%)	70 (34.0%)	268 (34.8%)	
genotype 1	462 (47.3%)	99 (48.1%)	363 (47.1%)	
genotype 2	160 (16.4%)	33 (16.0%)	127 (16.5%)	
Missing, n (%)	17 (1.7%)	4 (1.9%)	13 (1.7%)	
rs1179500 (C), n (%)				0.913
genotype 0	501 (51.3%)	107 (51.9%)	394 (51.1%)	
genotype 1	391 (40.0%)	82 (39.8%)	309 (40.1%)	
genotype 2	68 (7.0%)	13 (6.3%)	55 (7.1%)	
Missing, n (%)	17 (1.7%)	4 (1.9%)	13 (1.7%)	
rs12466981 (T), n (%)				0.232
genotype 0	528 (54.0%)	113 (54.9%)	415 (53.8%)	
genotype 1	367 (37.6%)	82 (39.8%)	285 (37.0%)	
genotype 2	74 (7.6%)	10 (4.9%)	64 (8.3%)	
Missing, n (%)	8 (0.8%)	1 (0.5%)	7 (0.9%)	
rs77972916 (A), n (%)				0.192
genotype 0	867 (88.7%)	183 (88.8%)	684 (88.7%)	
genotype 1	93 (9.5%)	20 (9.7%)	73 (9.5%)	
genotype 2	3 (0.3%)	2 (1.0%)	1 (0.1%)	
Missing, n (%)	14 (1.4%)	1 (0.5%)	13 (1.7%)	
rs17032590 (G), n (%)				0.237
genotype 0	605 (61.9%)	121 (58.7%)	484 (62.8%)	
genotype 1	301 (30.8%)	75 (36.4%)	226 (29.3%)	
genotype 2	38 (3.9%)	8 (3.9%)	30 (3.9%)	
Missing, n (%)	33 (3.4%)	2 (1.0%)	31 (4.0%)	
rs10173269 (G), n (%)				0.260
genotype 0	260 (26.6%)	55 (26.7%)	205 (26.6%)	
genotype 1	471 (48.2%)	93 (45.1%)	378 (49.0%)	
genotype 2	205 (21.0%)	52 (25.2%)	153 (19.8%)	
Missing, n (%)	41 (4.2%)	6 (2.9%)	35 (4.5%)	
rs11887136 (A), n (%)				0.897
genotype 0	746 (76.4%)	158 (76.7%)	588 (76.3%)	
genotype 1	204 (20.9%)	41 (19.9%)	163 (21.1%)	
genotype 2	13 (1.3%)	2 (1.0%)	11 (1.4%)	
Missing, n (%)	14 (1.4%)	5 (2.4%)	9 (1.2%)	
rs10184235 (G), n (%)				0.721
genotype 0	552 (56.5%)	120 (58.3%)	432 (56.0%)	
genotype 1	363 (37.2%)	76 (36.9%)	287 (37.2%)	
genotype 2	58 (5.9%)	10 (4.9%)	48 (6.2%)	
Missing, n (%)	4 (0.4%)	0 (0.0%)	4 (0.5%)	
rs207672 (T), n (%)				0.208
genotype 0	410 (42.0%)	96 (46.6%)	314 (40.7%)	

continued on next page

Table S.1 continued from previous page

Characteristic	Overall (N=977)	Alive (n=206)	Deceased (n=771)	*p*-value
genotype 1	426 (43.6%)	88 (42.7%)	338 (43.8%)	
genotype 2	113 (11.6%)	18 (8.7%)	95 (12.3%)	
Missing, n (%)	28 (2.9%)	4 (1.9%)	24 (3.1%)	
rs1561073 (T), n (%)				0.924
genotype 0	515 (52.7%)	112 (54.4%)	403 (52.3%)	
genotype 1	362 (37.1%)	75 (36.4%)	287 (37.2%)	
genotype 2	77 (7.9%)	17 (8.3%)	60 (7.8%)	
Missing, n (%)	23 (2.4%)	2 (1.0%)	21 (2.7%)	
rs9819463 (C), n (%)				0.835
genotype 0	612 (62.6%)	127 (61.7%)	485 (62.9%)	
genotype 1	318 (32.5%)	68 (33.0%)	250 (32.4%)	
genotype 2	45 (4.6%)	11 (5.3%)	34 (4.4%)	
Missing, n (%)	2 (0.2%)	0 (0.0%)	2 (0.3%)	
rs17821105 (T), n (%)				0.314
genotype 0	698 (71.4%)	139 (67.5%)	559 (72.5%)	
genotype 1	240 (24.6%)	58 (28.2%)	182 (23.6%)	
genotype 2	27 (2.8%)	7 (3.4%)	20 (2.6%)	
Missing, n (%)	12 (1.2%)	2 (1.0%)	10 (1.3%)	
rs113638840 (G), n (%)				0.352
genotype 0	520 (53.2%)	98 (47.6%)	422 (54.7%)	
genotype 1	362 (37.1%)	80 (38.8%)	282 (36.6%)	
genotype 2	52 (5.3%)	13 (6.3%)	39 (5.1%)	
Missing, n (%)	43 (4.4%)	15 (7.3%)	28 (3.6%)	
rs72811372 (A), n (%)				0.008
genotype 0	801 (82.0%)	162 (78.6%)	639 (82.9%)	
genotype 1	169 (17.3%)	40 (19.4%)	129 (16.7%)	
genotype 2	3 (0.3%)	3 (1.5%)	0 (0.0%)	
Missing, n (%)	4 (0.4%)	1 (0.5%)	3 (0.4%)	
rs7443323 (A), n (%)				0.350
genotype 0	537 (55.0%)	119 (57.8%)	418 (54.2%)	
genotype 1	365 (37.4%)	75 (36.4%)	290 (37.6%)	
genotype 2	68 (7.0%)	10 (4.9%)	58 (7.5%)	
Missing, n (%)	7 (0.7%)	2 (1.0%)	5 (0.6%)	
rs11745375 (T), n (%)				0.786
genotype 0	277 (28.4%)	56 (27.2%)	221 (28.7%)	
genotype 1	496 (50.8%)	109 (52.9%)	387 (50.2%)	
genotype 2	204 (20.9%)	41 (19.9%)	163 (21.1%)	
rs55993676 (T), n (%)				0.275
genotype 0	504 (51.6%)	112 (54.4%)	392 (50.8%)	
genotype 1	389 (39.8%)	82 (39.8%)	307 (39.8%)	
genotype 2	66 (6.8%)	9 (4.4%)	57 (7.4%)	
Missing, n (%)	18 (1.8%)	3 (1.5%)	15 (1.9%)	
rs196025 (A), n (%)				0.394
genotype 0	358 (36.6%)	68 (33.0%)	290 (37.6%)	
genotype 1	436 (44.6%)	94 (45.6%)	342 (44.4%)	
genotype 2	140 (14.3%)	34 (16.5%)	106 (13.7%)	
Missing, n (%)	43 (4.4%)	10 (4.9%)	33 (4.3%)	
rs9393688 (T), n (%)				0.051
genotype 0	495 (50.7%)	113 (54.9%)	382 (49.5%)	

continued on next page

Table S.1 continued from previous page

Characteristic	Overall (N=977)	Alive (n=206)	Deceased (n=771)	*p*-value
genotype 1	372 (38.1%)	65 (31.6%)	307 (39.8%)	
genotype 2	71 (7.3%)	20 (9.7%)	51 (6.6%)	
Missing, n (%)	39 (4.0%)	8 (3.9%)	31 (4.0%)	
rs58453446 (C), n (%)				0.563
genotype 0	414 (42.4%)	94 (45.6%)	320 (41.5%)	
genotype 1	451 (46.2%)	89 (43.2%)	362 (47.0%)	
genotype 2	110 (11.3%)	23 (11.2%)	87 (11.3%)	
Missing, n (%)	2 (0.2%)	0 (0.0%)	2 (0.3%)	
rs28719767 (C), n (%)				0.295
genotype 0	481 (49.2%)	111 (53.9%)	370 (48.0%)	
genotype 1	394 (40.3%)	74 (35.9%)	320 (41.5%)	
genotype 2	81 (8.3%)	18 (8.7%)	63 (8.2%)	
Missing, n (%)	21 (2.1%)	3 (1.5%)	18 (2.3%)	
rs28517513 (T), n (%)				0.970
genotype 0	568 (58.1%)	117 (56.8%)	451 (58.5%)	
genotype 1	362 (37.1%)	77 (37.4%)	285 (37.0%)	
genotype 2	43 (4.4%)	9 (4.4%)	34 (4.4%)	
Missing, n (%)	4 (0.4%)	3 (1.5%)	1 (0.1%)	
rs1528624 (G), n (%)				0.912
genotype 0	258 (26.4%)	54 (26.2%)	204 (26.5%)	
genotype 1	454 (46.5%)	93 (45.1%)	361 (46.8%)	
genotype 2	247 (25.3%)	54 (26.2%)	193 (25.0%)	
Missing, n (%)	18 (1.8%)	5 (2.4%)	13 (1.7%)	
rs1425794 (T), n (%)				0.747
genotype 0	330 (33.8%)	69 (33.5%)	261 (33.9%)	
genotype 1	465 (47.6%)	102 (49.5%)	363 (47.1%)	
genotype 2	182 (18.6%)	35 (17.0%)	147 (19.1%)	
rs17387279 (G), n (%)				0.446
genotype 0	654 (66.9%)	139 (67.5%)	515 (66.8%)	
genotype 1	271 (27.7%)	60 (29.1%)	211 (27.4%)	
genotype 2	38 (3.9%)	5 (2.4%)	33 (4.3%)	
Missing, n (%)	14 (1.4%)	2 (1.0%)	12 (1.6%)	
rs659398 (T), n (%)				0.247
genotype 0	499 (51.1%)	101 (49.0%)	398 (51.6%)	
genotype 1	368 (37.7%)	77 (37.4%)	291 (37.7%)	
genotype 2	73 (7.5%)	21 (10.2%)	52 (6.7%)	
Missing, n (%)	37 (3.8%)	7 (3.4%)	30 (3.9%)	
rs10987386 (T), n (%)				0.697
genotype 0	674 (69.0%)	146 (70.9%)	528 (68.5%)	
genotype 1	278 (28.5%)	54 (26.2%)	224 (29.1%)	
genotype 2	25 (2.6%)	6 (2.9%)	19 (2.5%)	
rs4948502 (C), n (%)				0.085
genotype 0	353 (36.1%)	83 (40.3%)	270 (35.0%)	
genotype 1	465 (47.6%)	84 (40.8%)	381 (49.4%)	
genotype 2	159 (16.3%)	39 (18.9%)	120 (15.6%)	
rs12571363 (A), n (%)				0.900
genotype 0	791 (81.0%)	165 (80.1%)	626 (81.2%)	
genotype 1	176 (18.0%)	39 (18.9%)	137 (17.8%)	
genotype 2	10 (1.0%)	2 (1.0%)	8 (1.0%)	

continued on next page

Table S.1 continued from previous page

Characteristic	Overall (N=977)	Alive (n=206)	Deceased (n=771)	*p*-value
rs7927422 (C), n (%)				0.410
genotype 0	349 (35.7%)	65 (31.6%)	284 (36.8%)	
genotype 1	445 (45.5%)	99 (48.1%)	346 (44.9%)	
genotype 2	161 (16.5%)	36 (17.5%)	125 (16.2%)	
Missing, n (%)	22 (2.3%)	6 (2.9%)	16 (2.1%)	
rs11022690 (C), n (%)				0.374
genotype 0	280 (28.7%)	53 (25.7%)	227 (29.4%)	
genotype 1	467 (47.8%)	100 (48.5%)	367 (47.6%)	
genotype 2	220 (22.5%)	53 (25.7%)	167 (21.7%)	
Missing, n (%)	10 (1.0%)	0 (0.0%)	10 (1.3%)	
rs11227223 (T), n (%)				0.107
genotype 0	871 (89.2%)	188 (91.3%)	683 (88.6%)	
genotype 1	96 (9.8%)	14 (6.8%)	82 (10.6%)	
genotype 2	2 (0.2%)	1 (0.5%)	1 (0.1%)	
Missing, n (%)	8 (0.8%)	3 (1.5%)	5 (0.6%)	
rs12313454 (G), n (%)				0.468
genotype 0	735 (75.2%)	152 (73.8%)	583 (75.6%)	
genotype 1	225 (23.0%)	52 (25.2%)	173 (22.4%)	
genotype 2	12 (1.2%)	1 (0.5%)	11 (1.4%)	
Missing, n (%)	5 (0.5%)	1 (0.5%)	4 (0.5%)	
rs10878300 (T), n (%)				0.819
genotype 0	694 (71.0%)	144 (69.9%)	550 (71.3%)	
genotype 1	247 (25.3%)	53 (25.7%)	194 (25.2%)	
genotype 2	36 (3.7%)	9 (4.4%)	27 (3.5%)	
rs4444235 (T), n (%)				0.200
genotype 0	256 (26.2%)	44 (21.4%)	212 (27.5%)	
genotype 1	486 (49.7%)	108 (52.4%)	378 (49.0%)	
genotype 2	235 (24.1%)	54 (26.2%)	181 (23.5%)	
rs1956028 (C), n (%)				0.942
genotype 0	739 (75.6%)	157 (76.2%)	582 (75.5%)	
genotype 1	222 (22.7%)	45 (21.8%)	177 (23.0%)	
genotype 2	14 (1.4%)	3 (1.5%)	11 (1.4%)	
Missing, n (%)	2 (0.2%)	1 (0.5%)	1 (0.1%)	
rs72743477 (G), n (%)				0.744
genotype 0	608 (62.2%)	124 (60.2%)	484 (62.8%)	
genotype 1	320 (32.8%)	72 (35.0%)	248 (32.2%)	
genotype 2	45 (4.6%)	9 (4.4%)	36 (4.7%)	
Missing, n (%)	4 (0.4%)	1 (0.5%)	3 (0.4%)	
rs4886509 (C), n (%)				0.677
genotype 0	422 (43.2%)	92 (44.7%)	330 (42.8%)	
genotype 1	417 (42.7%)	83 (40.3%)	334 (43.3%)	
genotype 2	103 (10.5%)	24 (11.7%)	79 (10.2%)	
Missing, n (%)	35 (3.6%)	7 (3.4%)	28 (3.6%)	
rs4889526 (A), n (%)				0.611
genotype 0	417 (42.7%)	94 (45.6%)	323 (41.9%)	
genotype 1	439 (44.9%)	87 (42.2%)	352 (45.7%)	
genotype 2	114 (11.7%)	25 (12.1%)	89 (11.5%)	
Missing, n (%)	7 (0.7%)	0 (0.0%)	7 (0.9%)	
rs7196853 (C), n (%)				0.683

continued on next page

Table S.1 continued from previous page

Characteristic	Overall (N=977)	Alive (n=206)	Deceased (n=771)	*p*-value
genotype 0	773 (79.1%)	157 (76.2%)	616 (79.9%)	
genotype 1	170 (17.4%)	40 (19.4%)	130 (16.9%)	
genotype 2	6 (0.6%)	1 (0.5%)	5 (0.6%)	
Missing, n (%)	28 (2.9%)	8 (3.9%)	20 (2.6%)	
rs750739 (G), n (%)				0.628
genotype 0	716 (73.3%)	151 (73.3%)	565 (73.3%)	
genotype 1	209 (21.4%)	41 (19.9%)	168 (21.8%)	
genotype 2	18 (1.8%)	5 (2.4%)	13 (1.7%)	
Missing, n (%)	34 (3.5%)	9 (4.4%)	25 (3.2%)	
rs4800410 (C), n (%)				0.299
genotype 0	358 (36.6%)	83 (40.3%)	275 (35.7%)	
genotype 1	444 (45.4%)	84 (40.8%)	360 (46.7%)	
genotype 2	132 (13.5%)	30 (14.6%)	102 (13.2%)	
Missing, n (%)	43 (4.4%)	9 (4.4%)	34 (4.4%)	
rs4636990 (G), n (%)				0.890
genotype 0	309 (31.6%)	68 (33.0%)	241 (31.3%)	
genotype 1	434 (44.4%)	90 (43.7%)	344 (44.6%)	
genotype 2	229 (23.4%)	47 (22.8%)	182 (23.6%)	
Missing, n (%)	5 (0.5%)	1 (0.5%)	4 (0.5%)	
rs1548029 (C), n (%)				0.687
genotype 0	383 (39.2%)	82 (39.8%)	301 (39.0%)	
genotype 1	453 (46.4%)	91 (44.2%)	362 (47.0%)	
genotype 2	141 (14.4%)	33 (16.0%)	108 (14.0%)	
rs72490631 (C), n (%)				0.091
genotype 0	592 (60.6%)	113 (54.9%)	479 (62.1%)	
genotype 1	331 (33.9%)	83 (40.3%)	248 (32.2%)	
genotype 2	49 (5.0%)	9 (4.4%)	40 (5.2%)	
Missing, n (%)	5 (0.5%)	1 (0.5%)	4 (0.5%)	
rs6006399 (G), n (%)				0.479
genotype 0	801 (82.0%)	175 (85.0%)	626 (81.2%)	
genotype 1	164 (16.8%)	29 (14.1%)	135 (17.5%)	
genotype 2	9 (0.9%)	2 (1.0%)	7 (0.9%)	
Missing, n (%)	3 (0.3%)	0 (0.0%)	3 (0.4%)	

Table S.2: Simulation results for the five non-zero covariates, averaged over the replicates indicated in column n_{valid}. Reported quantities are absolute bias (|Bias|), root mean squared error (RMSE), empirical standard error (EmpSE), the average model-based standard error (AvgSE) and the empirical coverage probability of the nominal 95% confidence intervals (Cov).

n	p	Method	n_{valid}	|Bias|	RMSE	EmpSE	AvgSE	Cov
500	20	Oracle	100	0.018	0.082	0.081	0.070	0.898
		SMC-DBL	100	0.013	0.088	0.088	0.084	0.938
		Standard IRO	100	0.087	0.121	0.083	0.075	0.748
		Std SMC-FCS	100	0.025	0.099	0.096	0.087	0.930
		Mean Imp + DL	100	0.052	0.096	0.080	0.071	0.840
		Complete Case + DL	100	0.029	0.127	0.124	0.109	0.918
	50	Oracle	100	0.025	0.083	0.078	0.069	0.916
		SMC-DBL	100	0.014	0.087	0.084	0.084	0.946

continued on next page

Table S.2 continued from previous page

n	p	Method	n_{valid}	|Bias|	RMSE	EmpSE	AvgSE	Cov
		Standard IRO	100	0.133	0.154	0.076	0.079	0.608
		Std SMC-FCS	100	0.054	0.118	0.104	0.091	0.884
		Mean Imp + DL	100	0.047	0.096	0.083	0.070	0.852
		Complete Case + DL	100	0.144	0.421	0.388	0.242	0.961
	100	Oracle	100	0.039	0.096	0.086	0.070	0.854
		SMC-DBL	100	0.031	0.099	0.093	0.085	0.898
		Standard IRO	100	0.184	0.201	0.081	0.090	0.470
		Std SMC-FCS	100	0.124	0.200	0.155	0.113	0.790
		Mean Imp + DL	100	0.041	0.100	0.088	0.071	0.836
		Complete Case + DL	–	–	–	–	–	–
	200	Oracle	100	0.063	0.108	0.083	0.071	0.802
		SMC-DBL	100	0.041	0.104	0.092	0.088	0.908
		Standard IRO	100	0.350	0.359	0.077	0.160	0.412
		Std SMC-FCS	14	5.354	5.639	1.680	2.580	0.471
		Mean Imp + DL	100	0.034	0.097	0.089	0.072	0.844
		Complete Case + DL	–	–	–	–	–	–
1000	20	Oracle	100	0.006	0.050	0.049	0.049	0.952
		SMC-DBL	100	0.017	0.060	0.057	0.059	0.950
		Standard IRO	100	0.061	0.083	0.056	0.051	0.722
		Std SMC-FCS	100	0.011	0.061	0.060	0.061	0.948
		Mean Imp + DL	100	0.067	0.087	0.054	0.050	0.712
		Complete Case + DL	100	0.010	0.084	0.083	0.077	0.924
	50	Oracle	100	0.012	0.055	0.054	0.049	0.926
		SMC-DBL	100	0.021	0.066	0.062	0.059	0.916
		Standard IRO	100	0.100	0.116	0.058	0.053	0.528
		Std SMC-FCS	100	0.021	0.074	0.071	0.062	0.902
		Mean Imp + DL	100	0.056	0.084	0.060	0.050	0.740
		Complete Case + DL	100	0.097	0.241	0.218	0.157	0.961
	100	Oracle	100	0.020	0.058	0.054	0.049	0.910
		SMC-DBL	100	0.019	0.064	0.060	0.059	0.926
		Standard IRO	100	0.129	0.140	0.054	0.056	0.366
		Std SMC-FCS	100	0.040	0.083	0.072	0.065	0.880
		Mean Imp + DL	100	0.053	0.078	0.056	0.050	0.754
		Complete Case + DL	–	–	–	–	–	–
	200	Oracle	100	0.034	0.066	0.055	0.049	0.850
		SMC-DBL	100	0.020	0.062	0.057	0.060	0.956
		Standard IRO	100	0.173	0.182	0.054	0.068	0.268
		Std SMC-FCS	100	0.103	0.139	0.091	0.078	0.752
		Mean Imp + DL	100	0.040	0.071	0.056	0.050	0.826
		Complete Case + DL	–	–	–	–	–	–
2000	20	Oracle	100	0.005	0.036	0.036	0.035	0.944
		SMC-DBL	100	0.015	0.043	0.040	0.042	0.948
		Standard IRO	100	0.045	0.060	0.039	0.036	0.728
		Std SMC-FCS	100	0.004	0.042	0.042	0.043	0.960
		Mean Imp + DL	100	0.071	0.081	0.037	0.036	0.490
		Complete Case + DL	100	0.014	0.059	0.057	0.055	0.948

continued on next page

Table S.2 continued from previous page

n	p	Method	n_{valid}	|Bias|	RMSE	EmpSE	AvgSE	Cov
	50	Oracle	100	0.010	0.040	0.039	0.035	0.918
		SMC-DBL	100	0.016	0.046	0.042	0.042	0.904
		Standard IRO	100	0.067	0.079	0.041	0.036	0.568
		Std SMC-FCS	100	0.013	0.048	0.046	0.043	0.932
		Mean Imp + DL	100	0.066	0.077	0.039	0.036	0.554
		Complete Case + DL	100	0.040	0.134	0.126	0.106	0.894
	100	Oracle	100	0.011	0.041	0.039	0.035	0.922
		SMC-DBL	100	0.017	0.048	0.044	0.042	0.912
		Standard IRO	100	0.090	0.099	0.040	0.037	0.374
		Std SMC-FCS	100	0.019	0.052	0.048	0.045	0.918
		Mean Imp + DL	100	0.059	0.071	0.039	0.036	0.598
		Complete Case + DL	100	0.403	0.539	0.349	0.061	0.112
	200	Oracle	100	0.025	0.048	0.040	0.035	0.852
		SMC-DBL	100	0.012	0.046	0.043	0.042	0.930
		Standard IRO	100	0.108	0.115	0.040	0.039	0.246
		Std SMC-FCS	100	0.046	0.069	0.050	0.046	0.800
		Mean Imp + DL	100	0.052	0.066	0.040	0.036	0.674
		Complete Case + DL	–	–	–	–	–	–

Table S.3: Model selection performance across simulation settings, averaged over the replicates indicated in column n_{valid}. Reported quantities are the average model size (MS), average number of true positives (TP), and average number of false positives (FP).

n	p	Method	n_{valid}	MS	TP	FP
500	20	Oracle	100	10.940	5.000	5.940
		SMC-DBL	100	9.760	5.000	4.760
		Standard IRO	100	10.320	5.000	5.320
		Std SMC-FCS	100	11.960	5.000	6.960
		Mean Imp + DL	100	11.380	5.000	6.380
		Complete Case + DL	100	9.990	5.000	4.990
	50	Oracle	100	15.560	5.000	10.560
		SMC-DBL	100	12.610	5.000	7.610
		Standard IRO	100	13.490	5.000	8.490
		Std SMC-FCS	100	22.590	5.000	17.590
		Mean Imp + DL	100	15.170	5.000	10.170
		Complete Case + DL	100	10.220	4.730	5.490
	100	Oracle	100	19.540	5.000	14.540
		SMC-DBL	100	13.590	5.000	8.590
		Standard IRO	100	15.120	5.000	10.120
		Std SMC-FCS	100	43.730	5.000	38.730
		Mean Imp + DL	100	20.130	5.000	15.130
		Complete Case + DL	–	–	–	–
	200	Oracle	100	21.900	5.000	16.900
		SMC-DBL	100	9.750	5.000	4.750
		Standard IRO	100	14.410	5.000	9.410
		Std SMC-FCS	14	200.000	5.000	195.000

continued on next page

Table S.3 continued from previous page

n	p	Method	n_{valid}	MS	TP	FP
		Mean Imp + DL	100	21.980	5.000	16.980
		Complete Case + DL	–	–	–	–
1000	20	Oracle	100	10.690	5.000	5.690
		SMC-DBL	100	10.260	5.000	5.260
		Standard IRO	100	10.930	5.000	5.930
		Std SMC-FCS	100	12.680	5.000	7.680
		Mean Imp + DL	100	11.430	5.000	6.430
		Complete Case + DL	100	10.390	5.000	5.390
	50	Oracle	100	16.500	5.000	11.500
		SMC-DBL	100	13.010	5.000	8.010
		Standard IRO	100	13.840	5.000	8.840
		Std SMC-FCS	100	22.480	5.000	17.480
		Mean Imp + DL	100	16.560	5.000	11.560
		Complete Case + DL	100	12.800	5.000	7.800
	100	Oracle	100	20.400	5.000	15.400
		SMC-DBL	100	15.230	5.000	10.230
		Standard IRO	100	16.180	5.000	11.180
		Std SMC-FCS	100	36.420	5.000	31.420
		Mean Imp + DL	100	19.840	5.000	14.840
		Complete Case + DL	100	2.929	1.111	1.818
	200	Oracle	100	25.660	5.000	20.660
		SMC-DBL	100	15.400	5.000	10.400
		Standard IRO	100	18.830	5.000	13.830
		Std SMC-FCS	100	74.830	5.000	69.830
		Mean Imp + DL	100	24.360	5.000	19.360
		Complete Case + DL	–	–	–	–
2000	20	Oracle	100	11.760	5.000	6.760
		SMC-DBL	100	10.950	5.000	5.950
		Standard IRO	100	11.260	5.000	6.260
		Std SMC-FCS	100	12.830	5.000	7.830
		Mean Imp + DL	100	11.950	5.000	6.950
		Complete Case + DL	100	11.120	5.000	6.120
	50	Oracle	100	17.170	5.000	12.170
		SMC-DBL	100	14.030	5.000	9.030
		Standard IRO	100	15.230	5.000	10.230
		Std SMC-FCS	100	24.250	5.000	19.250
		Mean Imp + DL	100	17.910	5.000	12.910
		Complete Case + DL	100	13.490	5.000	8.490
	100	Oracle	100	19.880	5.000	14.880
		SMC-DBL	100	15.710	5.000	10.710
		Standard IRO	100	16.850	5.000	11.850
		Std SMC-FCS	100	36.820	5.000	31.820
		Mean Imp + DL	100	20.460	5.000	15.460
		Complete Case + DL	100	6.580	2.860	3.720
	200	Oracle	100	26.730	5.000	21.730

continued on next page

Table S.3 continued from previous page

n	p	Method	n_{valid}	MS	TP	FP
		SMC-DBL	100	18.370	5.000	13.370
		Standard IRO	100	19.950	5.000	14.950
		Std SMC-FCS	100	64.206	5.000	59.206
		Mean Imp + DL	100	25.670	5.000	20.670
		Complete Case + DL	–	–	–	–

.