Biostatistics (2021) **22**, 2, *pp.* 381–401
doi:10.1093/biostatistics/kxz036
Advance Access publication on September 23, 2019

A fast divide-and-conquer sparse Cox regression

YAN WANG

Department of Environmental Health, Harvard T. H. Chan School of Public Health, 401 Park Drive West, Boston, MA, 02215, USA and Department of Biostatistics, Harvard T. H. Chan School of Public Health, 655 Huntington Avenue, Boston, MA, 02115, USA

CHUAN HONG, NATHAN PALMER

Department of Biomedical Informatics, Harvard Medical School, 25 Shattuck Street, Boston, MA, 02115, USA

QIAN DI, JOEL SCHWARTZ

Department of Environmental Health, Harvard T. H. Chan School of Public Health, 401 Park Drive West, Boston, MA, 02215, USA

ISAAC KOHANE

Department of Biomedical Informatics, Harvard Medical School, 25 Shattuck Street, Boston, MA, 02115, USA

TIANXI CAI*

Department of Biostatistics, Harvard T. H. Chan School of Public Health, 655 Huntington Avenue, Boston, MA, 02115, USA and Department of Biomedical Informatics, Harvard Medical School, 25 Shattuck Street, Boston, MA, 02115, USA

tcai@hsph.harvard.edu

Summary

We propose a computationally and statistically efficient divide-and-conquer (DAC) algorithm to fit sparse Cox regression to massive datasets where the sample size n_0 is exceedingly large and the covariate dimension p is not small but $n_0 \gg p$. The proposed algorithm achieves computational efficiency through a one-step linear approximation followed by a least square approximation to the partial likelihood (PL). These sequences of linearization enable us to maximize the PL with only a small subset and perform penalized estimation via a fast approximation to the PL. The algorithm is applicable for the analysis of both time-independent and time-dependent survival data. Simulations suggest that the proposed DAC algorithm substantially outperforms the full sample-based estimators and the existing DAC algorithm with respect to the computational speed, while it achieves similar statistical efficiency as the full sample-based estimators. The proposed algorithm was applied to extraordinarily large survival datasets for the prediction of heart failure-specific readmission within 30 days among Medicare heart failure patients.

Keywords: Cox proportional hazards model; Distributed learning; Divide-and-conquer; Least square approximation; Shrinkage estimation; Variable selection.

*To whom correspondence should be addressed.

1. INTRODUCTION

Large datasets derived from health insurance claims and electronic health records are increasingly available for healthcare and medical research. These datasets are valuable sources for the development of risk prediction models, which are the key components of precision medicine. Fitting risk prediction models to a dataset with a massive sample size (n_0), however, is computationally challenging, especially when the number of candidate predictors (p) is also large and yet only a small subset of the predictors is informative. While it is statistically feasible to fit a full model with all p variables, deriving parsimonious risk prediction models has the advantage of being more clinically interpretable and easier to implement in practice. In such a setting, it is highly desirable to fit a sparse regression model to simultaneously remove non-informative predictors and estimate the effects of the informative predictors. When the outcome of interest is time-to-event and is subject to censoring, one may obtain a sparse risk prediction model by fitting a regularized Cox proportional hazards model (Cox, 1972) with penalty functions such as the adaptive least absolute shrinkage and selection operator (LASSO) penalty (Zhang and Lu, 2007).

When n_0 is extraordinarily large, directly fitting an adaptive LASSO (aLASSO) penalized Cox model to such a dataset is not computationally feasible. To overcome the computational difficulty, one may employ the divide-and-conquer (DAC) strategy, which typically divides the full sample into K subsets, solves the optimization problem using each subset, and combines the subset-specific estimates into a combined estimate. Various DAC algorithms have been proposed to fit penalized regression models. For example, Chen and Xie (2014) proposed a DAC algorithm to fit penalized generalized linear models (GLMs). The algorithm obtains a sparse GLM estimate for each subset and then combine subset-specific estimates by majority voting and averaging. Tang *and others* (2016) proposed an alternative DAC algorithm to fit GLM with an extremely large n_0 and a large p by combining de-biased LASSO estimates from each subset. While both algorithms are effective in reducing the computation burden compared to fitting a penalized regression model to the full data, they remain computationally intensive as K penalized estimation procedures will be required. In addition, the DAC strategy has not been extended to the survival data analysis.

In this article, we propose a novel DAC algorithm using a sequence of linearizations, denoted by DAC_{lin}, to fit aLASSO penalized Cox proportional hazards models, which can further reduce the computation burden compared to the existing DAC algorithms. DAC_{lin} starts with obtaining an estimator that maximizes the partial likelihood (PL) of a subset of the full data, which is then updated using all subsets via one-step approximations. The updated estimator serves as a $\sqrt{n_0}$-consistent initial estimator for the aLASSO problem and approximates the full sample-based maximum PL estimator. Subsequently, we obtain the final aLASSO estimator based on an objective function applying the least square approximation (LSA) to the PL as in Wang and Leng (2007). The LSA allows us to fit the aLASSO using a pseudo-likelihood based on a sample of size p. The penalized regression is only fit once to a $p \times p$ dimensional psuedo data in the proposed DAC_{lin} algorithm and the improvement in computation cost is substantial if $n_0 \gg p$. Our proposed DAC_{lin} algorithm can also accommodate time-dependent covariates.

The rest of the article is organized as follows. We detail the DAC_{lin} algorithm in Section 2. In Section 3, we present simulation results demonstrating the superiority of DAC_{lin} compared to the existing methods when covariates are time-independent and when some covariates are time-dependent. In Section 4, we employ the DAC_{lin} algorithm to develop risk prediction models for 30-day readmission after an index heart failure hospitalization with data from over 10 million Medicare patients by fitting regularized Cox models with (i) $p = 540$ time-independent covariates and (ii) $p_{\text{ind}} = 575$ time-independent covariates and $p_{\text{dep}} = 5$ time-dependent environmental covariates. We conclude with some discussions in Section 5.

2. Methods

2.1. *Notation and settings*

Let T denote the survival time and $\mathbf{Z}(\cdot)$ denote the $p \times 1$ vector of bounded and potentially time-dependent covariates. Due to censoring, for T, we only observe (X, Δ), where $X = \min(T, C)$, $\Delta = I(T \leq C)$, and C is the censoring time assumed to be independent of T given $\mathbf{Z}(\cdot)$. Suppose the data for analysis consist of n_0 subjects with independent realizations of $\mathbf{D} = (X, \Delta, \mathbf{Z}(\cdot)^\mathsf{T})^\mathsf{T}$, denoted by $\mathscr{D}_{\text{full}} = \{\mathbf{D}_i = (X_i, \Delta_i, \mathbf{Z}_i(\cdot)^\mathsf{T})^\mathsf{T}, i = 1, ..., n_0\}$, where we assume that $n_0 \gg p$.

We denote the index set for the full data by $\Omega_{\text{full}} = \{1, ..., n_0\}$. For all DAC algorithms discussed in this article, we randomly partition $\mathscr{D}_{\text{full}}$ into K subsets with the k-th subset denoted by $\mathscr{D}_k = \{\mathbf{D}_i, i \in \Omega_k\}$. Without loss of generality, we assume that $n = n_0/K$ is an integer and that the index set for the subset k is $\Omega_k = \{(k-1)n + 1, ..., kn\}$. For any index set Ω, we denote the size of Ω by n_Ω with $n_\Omega = n_0$ if $\Omega = \Omega_{\text{full}}$ and $n_\Omega = n$ if $\Omega = \Omega_k$. Throughout we assume that $K = o\left(n_0^{\frac{1}{2}}\right)$ such that $n^{-1} = o\left(n_0^{-\frac{1}{2}}\right)$ and $n \gg p$.

We aim to predict T based on $\mathbf{Z}(\cdot)$ via a Cox model with conditional hazard function

$$\lambda(t|\mathbf{Z}(t)) = \lambda_0(t) \exp\left(\beta_0^\mathsf{T}\mathbf{Z}(t)\right), \tag{2.1}$$

where $\lambda_0(t)$ is the baseline hazard function. Our goal is to develop a computationally and statistically efficient procedure to estimate $\boldsymbol{\beta}_0$ using $\mathscr{D}_{\text{full}}$ under the assumption that $\boldsymbol{\beta}_0$ is sparse with the size of the active set $\mathcal{A} = \{\jmath : \beta_{0\jmath} \neq 0\}$ much smaller than p. When n_0 is not extraordinarily large, we may obtain an efficient estimate, denoted by $\widehat{\boldsymbol{\beta}}_{\text{full}}$, based on the aLASSO penalized PL likelihood estimator as proposed in Zhang and Lu (2007). Specifically,

$$\widehat{\boldsymbol{\beta}}_{\text{full}} = \underset{\boldsymbol{\beta}}{\text{argmax}} \left\{ \widehat{\ell}_{\Omega_{\text{full}}}(\boldsymbol{\beta}) - \lambda_{\Omega_{\text{full}}} \sum_{\iota=1}^{p} \frac{|\beta_\iota|}{|\widetilde{\beta}_{\iota,\text{init}}|^\gamma} \right\} \tag{2.2}$$

where for any index set Ω,

$$\widehat{\ell}_\Omega(\boldsymbol{\beta}) = n_\Omega^{-1} \sum_{i \in \Omega} \ell_i(\boldsymbol{\beta}), \quad \ell_i(\boldsymbol{\beta}) = \Delta_i \left[\boldsymbol{\beta}^\mathsf{T}\mathbf{Z}_i(X_i) - \log\left\{ \sum_{i' \in \Omega} I(X_{i'} \geq X_i) e^{\boldsymbol{\beta}^\mathsf{T}\mathbf{Z}_{i'}(X_{i'})} \right\} \right], \tag{2.3}$$

$\widetilde{\boldsymbol{\beta}}_{\text{init}} = (\widetilde{\beta}_{1,\text{init}}, \cdots, \widetilde{\beta}_{p,\text{init}})'$ is an initial $\sqrt{n_0}$-consistent estimator of model (2.1), $\lambda_{\Omega_{\text{full}}} \geq 0$ is a tuning parameter, and $\gamma > 0$. A simple choice of $\widetilde{\boldsymbol{\beta}}_{\text{init}}$ is $\widetilde{\boldsymbol{\beta}}_{\Omega_{\text{full}}}$, where for any set Ω,

$$\widetilde{\boldsymbol{\beta}}_\Omega = \underset{\boldsymbol{\beta}}{\text{argmax}}\, \widehat{\ell}_\Omega(\boldsymbol{\beta}).$$

Following the arguments given in Zhang and Lu (2007), when $n_0^{\frac{1}{2}} \lambda_{\Omega_{\text{full}}} \to 0$ and $n_0^{(1+\gamma)/2} \lambda_{\Omega_{\text{full}}} \to \infty$, we can show that $\widehat{\boldsymbol{\beta}}_{\text{full}}$ achieves the variable selection consistency, i.e. the estimated active set $\widehat{\mathcal{A}}_{\text{full}} = \{\jmath : \widehat{\beta}_{\text{full},\jmath} \neq 0\}$ satisfies $P(\widehat{\mathcal{A}}_{\text{full}} = \mathcal{A}) \to 1$ and that the oracle property holds, i.e.

$$n_0^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{\text{full}}^{\mathcal{A}} - \boldsymbol{\beta}_0^{\mathcal{A}}) = \widehat{\mathbb{A}}_{\Omega_{\text{full}}}^{\mathcal{A}}(\boldsymbol{\beta}_0)^{-1} n_0^{\frac{1}{2}} \widehat{\mathbf{U}}_{\Omega_{\text{full}}}^{\mathcal{A}}(\boldsymbol{\beta}_0) + o_p(1) = \mathbb{A}^{\mathcal{A}}(\boldsymbol{\beta}_0)^{-1} n_0^{\frac{1}{2}} \widehat{\mathbf{U}}_{\Omega_{\text{full}}}^{\mathcal{A}}(\boldsymbol{\beta}_0) + o_p(1) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \mathbb{A}^{\mathcal{A}}(\boldsymbol{\beta}_0)^{-1}\right),$$

where $\mathbf{G}^{\mathcal{A}} = \{W_l, l \in \mathcal{A}\}$ if $\mathbf{G}$ is a vector and $\mathbf{G}^{\mathcal{A}} = [W_{ll'}]_{l\in\mathcal{A},l'\in\mathcal{A}}$ if $\mathbf{G}$ is a matrix,

$$\widehat{\mathbf{U}}_\Omega(\boldsymbol{\beta}) = n_\Omega^{-1}\sum_{i\in\Omega}\widehat{\mathbf{U}}_{i,\Omega}(\boldsymbol{\beta}), \quad \widehat{\mathbf{U}}_{i,\Omega}(\boldsymbol{\beta}) = \int\{\mathbf{Z}_i(t) - \widehat{\mathcal{S}}_{1,\Omega}(t,\boldsymbol{\beta})/\widehat{\mathcal{S}}_{0,\Omega}(t,\boldsymbol{\beta})\}\mathrm{d}M_i(t,\boldsymbol{\beta}),$$

$$\widehat{\mathcal{S}}_{r,\Omega}(t,\boldsymbol{\beta}) = n_\Omega^{-1}\sum_{i\in\Omega}\mathbf{Z}_i(t)^{\otimes r}I(X_i \geq t), \quad \widehat{\mathbb{A}}_\Omega(\boldsymbol{\beta}) = -\partial^2\widehat{\ell}_\Omega(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathsf{T}},$$

$$\mathbb{A}(\boldsymbol{\beta}) = \int\frac{\mathcal{S}_2(t,\boldsymbol{\beta})\mathcal{S}_0(t,\boldsymbol{\beta}) - \mathcal{S}_1(t,\boldsymbol{\beta})^{\otimes 2}}{\mathcal{S}_0(t,\boldsymbol{\beta})^2}\mathrm{d}E\{N_i(t)\}, \quad \mathcal{S}_r(t,\boldsymbol{\beta}) = E\{\mathbf{Z}_i(t)^{\otimes r}I(X_i \geq t)\},$$

$N_i(t) = I(T_i \leq t)\Delta_i$, $M_i(t,\boldsymbol{\beta}) = N_i(t) - \int_0^t I(X_i \geq u)\mathrm{e}^{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}_i(u)}\lambda_0(u)\mathrm{d}u$, and for any vector $\mathbf{a}$, $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathsf{T}}$.

When n_0 is not too large, multiple algorithms are available to solve (2.2) with time-independent covariates, including a coordinate gradient descent algorithm (Simon *and others*, 2011), a least angle regression-like algorithm (Park and Hastie, 2007), a combination of gradient descent-Newton Raphson method (Goeman, 2010), and a modified shooting algorithm (Zhang and Lu, 2007). Unfortunately, when n_0 is extraordinarily large, existing algorithms for fitting (2.2) are highly computationally intensive and subject to memory constraints. These algorithms may even be infeasible to compute in the presence of time-dependent covariates as each subject contribute multiple observations in the fitting.

2.2. *The DAC$_{\mathsf{lin}}$ algorithm*

The goal of this article is to develop an estimator that achieves the same asymptotic efficiency as $\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$ but can be computed very efficiently.

Our proposed algorithm, DAC$_{\mathsf{lin}}$, for attaining such a property is motivated by the LSA proposed in Wang and Leng (2007), with the LSA applied to the full sample-based PL. Specifically, it is not difficult to show that $\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$ is asymptotically equivalent to $\widehat{\boldsymbol{\beta}}_{\mathsf{full,lin}}$, where

$$\widehat{\boldsymbol{\beta}}_{\mathsf{full,lin}} = \underset{\boldsymbol{\beta}}{\mathrm{argmin}}\,\frac{1}{2}(\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}} - \boldsymbol{\beta})^{\mathsf{T}}\widehat{\mathbb{A}}_{\Omega_{\mathsf{full}}}(\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}})(\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}} - \boldsymbol{\beta}) + \lambda_{n_0}\sum_{\iota=1}^{p}\frac{|\beta_\iota|}{|\widetilde{\beta}_{\iota,\Omega_{\mathsf{full}}}|^\gamma}$$

That is, $\widehat{\boldsymbol{\beta}}_{\mathsf{full,lin}}$ will also achieve the variable selection consistency as $\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$ and $\widehat{\boldsymbol{\beta}}^{\mathcal{A}}_{\mathsf{full,lin}}$ has the same limiting distribution as that of $\widehat{\boldsymbol{\beta}}^{\mathcal{A}}_{\mathsf{full}}$. This suggests that an estimator can recover the distribution of $\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$ if we can construct accurate DAC approximations to $\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$ and $\widehat{\mathbb{A}}_{\Omega_{\mathsf{full}}}(\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}})$. To this end, we propose a linearization-based DAC estimator, denoted by $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$, which requires three main steps: (i) obtaining an estimator for the unpenalized problem $\widetilde{\boldsymbol{\beta}}^{[0]}_{\mathsf{DAC}}$ based on a subset, say $\mathscr{D}_1$; (ii) obtaining updated estimators for the unpenalized problem through one-step approximations using all K subsets; and (iii) constructing an aLASSO penalized estimator based on LSA. The procedure also brings a $\widehat{\mathbb{A}}_{\mathsf{DAC}}(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}})$ that well approximates $\widehat{\mathbb{A}}_{\Omega_{\mathsf{full}}}(\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}})$.

Specifically, in step (i), we use subset $\mathscr{D}_1$ to obtain a standard maximum PL estimator,

step (i) $$\widetilde{\boldsymbol{\beta}}^{[0]}_{\mathsf{DAC}} \equiv \widetilde{\boldsymbol{\beta}}_{\Omega_1} = \underset{\boldsymbol{\beta}}{\mathrm{argmax}}\,\widehat{\ell}_{\Omega_1}(\boldsymbol{\beta}).$$

In step (ii), we obtain a DAC one-step approximation to $\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$,

step (ii) $$\text{for } \iota = 1, \ldots, \mathcal{I}, \quad \widetilde{\boldsymbol{\beta}}^{[\iota]}_{\mathsf{DAC}} = K^{-1}\sum_{k=1}^{K}\widetilde{\boldsymbol{\beta}}_{\Omega_{k,\mathsf{lin}}}(\widetilde{\boldsymbol{\beta}}^{[\iota-1]}_{\mathsf{DAC}})$$

where

$$\widetilde{\boldsymbol{\beta}}_{\Omega_{k,\mathsf{lin}}}(\boldsymbol{\beta}) = \boldsymbol{\beta} + \widehat{\mathbb{A}}_{\mathsf{DAC}}(\boldsymbol{\beta})^{-1}\widehat{\mathbf{U}}_{\Omega_k}(\boldsymbol{\beta}) \quad \text{and} \quad \widehat{\mathbb{A}}_{\mathsf{DAC}}(\boldsymbol{\beta}) = K^{-1}\sum_{k=1}^{K}\widehat{\mathbb{A}}_{\Omega_k}(\boldsymbol{\beta}). \tag{2.4}$$

Let $\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}} = \widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}^{[\mathcal{I}]}$ be our DAC approximation to $\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$. In practice, we find that it suffices to let $\mathcal{I} = 2$. Finally, we apply the LSA to the PL and approximate $\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$ using $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$, where

$$\text{step (iii)} \qquad \widehat{\boldsymbol{\beta}}_{\mathsf{DAC}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\frac{1}{2}(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}} - \boldsymbol{\beta})^{\mathsf{T}}\widehat{\mathbb{A}}_{\mathsf{DAC}}(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}})(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}} - \boldsymbol{\beta}) + \lambda_{\Omega_{\mathsf{full}}}\sum_{J=1}^{p}\frac{|\beta_J|}{|\widetilde{\beta}_{\mathsf{DAC},J}|^{\gamma}}\right\}.$$

The optimization problem in step (iii) is equivalent to

$$\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\frac{1}{2}(\widetilde{\mathbf{Y}}_0(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}) - \widetilde{\mathbf{X}}_0(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}})\boldsymbol{\beta})^{\mathsf{T}}(\widetilde{\mathbf{Y}}_0(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}) - \widetilde{\mathbf{X}}_0(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}})\boldsymbol{\beta}) + \lambda_{\Omega_{\mathsf{full}}}\sum_{J=1}^{p}\frac{|\beta_J|}{|\widetilde{\beta}_{\mathsf{DAC},J}|^{\gamma}}\right\} \tag{2.5}$$

where $\tilde{\mathbf{Y}}_0(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}) = \widehat{\mathbb{A}}_{\mathsf{DAC}}(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}})^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$ is a $p\times 1$ vector and $\tilde{\mathbf{X}}_0(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}) = \widehat{\mathbb{A}}_{\mathsf{DAC}}(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}})^{\frac{1}{2}}$ is a $p\times p$ matrix. The linearization in step (iii) is exactly the same as that in Zhang and Lu (2007), which allows us to solve the penalized regression step using a pseudo likelihood based on a sample of size p. The computation cost of this step compared to solving (2.2) reduces substantially when $n_0 \gg p$. In the Appendix of the supplementary material available at *Biostatistics* online, we show that $n_0^{\frac{1}{2}}(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}} - \widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}) = o_p(1)$. It then follows from the similar arguments given in Wang and Leng (2007) that if $n_0^{\frac{1}{2}}\lambda_{n_0} \to 0$, $n_0^{(1+\gamma)/2}\lambda_{n_0} \to \infty$, the estimated active set using $\mathsf{DAC}_{\mathsf{lin}}$ $\widehat{\mathcal{A}}_{\mathsf{DAC}}$ achieves the variable selection consistency, i.e. $P(\widehat{\mathcal{A}}_{\mathsf{DAC}} = \mathcal{A}) \to 1$ and the oracle property holds, i.e. $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}^{\mathcal{A}}$ and $\widehat{\boldsymbol{\beta}}_{\mathsf{full}}^{\mathcal{A}}$ have the same limiting distribution.

2.3. *Tuning and standard error calculation*

The tuning parameter $\lambda_{\Omega_{\mathsf{full}}}$ is chosen by minimizing the Bayesian information criteria (BIC) of the fitted model. Volinsky and Raftery (2000) showed that the exact Bayes factor can be better approximated for the Cox model if the number of uncensored cases, $d_0 = \sum_{i\in\Omega_{\mathsf{full}}}\Delta_i$, is used to penalize the degrees of freedom in the BIC. Specifically, for any given tuning parameter $\lambda_{\Omega_{\mathsf{full}}}$ with its corresponding estimate of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\lambda_{\Omega_{\mathsf{full}}}}$, the BIC suggested by Volinsky and Raftery (2000) is defined as

$$\mathrm{BIC}_{V,\lambda_{\Omega_{\mathsf{full}}}} = -2\sum_{i\in\Omega_{\mathsf{full}}}\ell_i(\widehat{\boldsymbol{\beta}}_{\lambda_{\Omega_{\mathsf{full}}}}) + (\log d_0)\mathrm{d}f_{\lambda_{\Omega_{\mathsf{full}}}}, \tag{2.6}$$

where $df_{\lambda_{\Omega_{\mathsf{full}}}} = \sum_{J=1}^{p} I(\widehat{\beta}_{\lambda_{\Omega_{\mathsf{full}}},J} \neq 0)$. With the LSA, we may further approximate $\mathrm{BIC}_{V,\lambda_{\Omega_{\mathsf{full}}}}$ by

$$\mathrm{BIC}_{V_L,\lambda_{\Omega_{\mathsf{full}}}} = n_0(\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}} - \widehat{\boldsymbol{\beta}}_{\lambda_{\Omega_{\mathsf{full}}}})^{\mathsf{T}}\widehat{\mathbb{A}}_{\mathsf{DAC}}(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}})(\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}} - \widehat{\boldsymbol{\beta}}_{\lambda_{\Omega_{\mathsf{full}}}}) + (\log d_0)df_{\lambda_{\Omega_{\mathsf{full}}}}. \tag{2.7}$$

For the estimation of $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$, we chose a $\lambda_{\Omega_{\mathsf{full}}}$ such that $\mathrm{BIC}_{V_L,\lambda_{\Omega_{\mathsf{full}}}}$ is minimized. The oracle property is expected to hold in the setting where $n_0 \gg p$ and n_0 is extraordinarily large. We may thus estimate the variance-covariance matrix for $n_0^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}^{\mathcal{A}} - \boldsymbol{\beta}_0^{\mathcal{A}})$ using $\{\widehat{\mathbb{A}}_{\mathsf{DAC}}^{\mathcal{A}}(\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}})\}^{-1}$. For $J \in \mathcal{A}$, a $(1-\alpha)\times 100\%$ confidence interval for β_{0J} can be calculated accordingly.

3. SIMULATIONS

3.1. *Simulation settings*

We performed two sets of simulations to evaluate the performance of $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$ for the fitting of sparse Cox models, one with only time-independent covariates and the other with time-dependent covariates. For both scenarios, we focused primarily on $n_0 = 10^6$ and $K = 100$. We consider the number of iterations $\mathcal{I} = 1, 2$, and 3 to examine the impact of $\mathcal{I}$ on the proposed estimator.

3.1.1. *Time-independent covariates* We conducted extensive simulations to evaluate the performance of the proposed estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$ relative to (a) the performance of the full sample-based aLASSO estimator for the Cox model $\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$ and (b) a majority voting-based DAC method for the Cox model, denoted by $\widehat{\boldsymbol{\beta}}_{\mathsf{MV}}$ also with $K = 100$, penalized by a minimax concave penalty (MCP), which extends the majority voting-based DAC scheme for GLM proposed by Chen and Xie (2014). The reason of choosing $\widehat{\boldsymbol{\beta}}_{\mathsf{MV}}$ as a comparison is that there is no other DAC method available for the Cox model and only Chen and Xie (2014) considered a similar majority voting-based DAC method for the penalized GLM with non-adaptive penalties. We set a *priori* that $\widehat{\boldsymbol{\beta}}_{\mathsf{MV}}$ sets the estimate of a coefficient at zero, if at least 50% of the subset-specific estimates have a zero estimate for that coefficient. In addition, we compared the performance of the DAC estimator $\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$ relative to the full sample maximum PL estimator $\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$.

For the penalized procedures, we selected the tuning parameter based on the BIC discussed in Section 2.3. The aLASSO procedures were fit using the `glmnet` function in R package `glmnet` (Friedman *and others*, 2010; Simon *and others*, 2011; R Core Team, 2017) with $\gamma = 1$; the MCP procedures were fit using the `ncvsurv` function in R package `ncvreg` (Breheny and Huang, 2011; R Core Team, 2017). When there are ties among failure times, we used the Efron's method within each data subset (Efron, 1977).

For the covariates, we considered $p = 50$ and $p = 200$. We generated $\mathbf{Z}$ from a multivariate normal distribution with mean $\mathbf{0}_p^{\mathsf{T}}$ and variance-covariance matrix $\mathbb{V} = [I(l = l') + \nu I(l \neq l')]_{l=1,\ldots,p}^{l'=1,\ldots,p}$, where $\mathbf{a}_q$ denotes a $q \times 1$ vector with all elements being a and we considered $\nu = 0.2$, 0.5, and 0.8 to represent weak, moderate, and strong correlations among the covariates. For a given $\mathbf{Z}_i, i = 1, \cdots, n_0$, we generated T_i from a Weibull distribution with a shape parameter of 2 and a scale parameter of $\{0.5 \exp(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{Z}_i)\}^{-0.5}$, where we considered three choices of $\boldsymbol{\beta}_0$ to reflect different degrees of sparsity and signal strength:

$$\begin{aligned}
\boldsymbol{\beta}_0^{(\mathrm{I})} &= (\mathbf{0.8}_3^{\mathsf{T}}, \mathbf{0.4}_3^{\mathsf{T}}, \mathbf{0.2}_3^{\mathsf{T}}, \mathbf{0}_{p-9}^{\mathsf{T}})^{\mathsf{T}}, \\
\boldsymbol{\beta}_0^{(\mathrm{II})} &= (\mathbf{0.4}_4^{\mathsf{T}}, \mathbf{0.2}_4^{\mathsf{T}}, \mathbf{0.1}_4^{\mathsf{T}}, \mathbf{0.05}_4^{\mathsf{T}}, \mathbf{0}_{p-9}^{\mathsf{T}})^{\mathsf{T}}, \quad \text{and} \\
\boldsymbol{\beta}_0^{(\mathrm{III})} &= (1, 0.5, \mathbf{0.2}_2^{\mathsf{T}}, \mathbf{0.1}_2^{\mathsf{T}}, \mathbf{0.05}_2^{\mathsf{T}}, \mathbf{0.035}_3^{\mathsf{T}}, \mathbf{0}_{p-11})^{\mathsf{T}}.
\end{aligned}$$

For censoring, we generated C from an exponential distribution with a rate parameter of exp(0.5), resulting in 68% $\sim$ 76% of censoring across different configurations.

We additionally considered $n_0 = 10^5$, 10^6, 2×10^6 and 9×10^6 to evaluate how different choices of n_0 impact the relative performance of different procedures.

3.1.2. *Time-dependent covariates* We also conducted simulations for the settings where time-dependent covariates are present to evaluate the performance of $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$. Since neither `glmnet` nor `ncvsurv` allows time-dependent survival data, we used $\widehat{\boldsymbol{\beta}}_{\mathsf{full,lin}}$ as a benchmark to compare $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$ with. In addition, we compared the performance of $\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$ relative to $\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$.

We considered $p = 100$ consisting of $p_{\mathsf{ind}} = 50$ time-independent covariates and $p_{\mathsf{dep}} = 50$ time-dependent covariates. The simulation of the survival data with time-dependent covariates extended the

simulation scheme of Austin (2012) from dichotomous time-dependent covariates to continuous time-dependent covariates. We considered four time intervals $R_1 = [0,1)$, $R_2 = [1,2)$, $R_3 = [2,3)$, and $R_4 = [3,\infty)$, where the time-dependent covariates are constant within each interval but can vary between intervals. We generated $\mathbf{Z} = (\mathbf{Z}_{\text{ind}}^{\mathsf{T}}, \mathbf{Z}_{\text{dep}}(t \in R_1)^{\mathsf{T}}, \mathbf{Z}_{\text{dep}}(t \in R_2)^{\mathsf{T}}, \mathbf{Z}_{\text{dep}}(t \in R_3)^{\mathsf{T}}, \mathbf{Z}_{\text{dep}}(t \in R_4)^{\mathsf{T}})^{\mathsf{T}}$ from a multivariate normal distribution with mean $\mathbf{0}_{p_{\text{ind}}+4p_{\text{dep}}}^{\mathsf{T}}$ and variance-covariance matrix $\mathbb{V} = [I(l = l') + vI(l \neq l')]_{l=1,\ldots,p_{\text{ind}}+4p_{\text{dep}}}^{l'=1,\ldots,p_{\text{ind}}+4p_{\text{dep}}}$, where $\mathbf{Z}_{\text{ind}}$ are the time-independent covariates and $\mathbf{Z}_{\text{dep}}(t \in R_J)$ are the time-dependent for $t \in R_J$. We similarly considered $v = 0.2, 0.5, 0.8$ to represent weak, moderate, and strong correlations.

We generated T_i from a Weibull distribution with a shape parameter of 2 and a scale parameter of $\{0.05 \exp(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{Z}_i(t))\}^{-0.5}$, where $\boldsymbol{\beta}_0^{(\text{IV})} = (\boldsymbol{\beta}_{\text{ind},0}^{(\text{IV})^{\mathsf{T}}}, \boldsymbol{\beta}_{\text{dep},0}^{(\text{IV})^{\mathsf{T}}})^{\mathsf{T}}$,

$$\boldsymbol{\beta}_{\text{ind},0}^{(\text{IV})} = (\mathbf{0.08}_3^{\mathsf{T}}, \mathbf{0.04}_3^{\mathsf{T}}, \mathbf{0.02}_3^{\mathsf{T}}, \mathbf{0}_{p_{\text{ind}}-9}^{\mathsf{T}})^{\mathsf{T}}, \text{ and } \quad \boldsymbol{\beta}_{\text{dep},0}^{(\text{IV})} = (\mathbf{0.08}_3^{\mathsf{T}}, \mathbf{0.04}_3^{\mathsf{T}}, \mathbf{0.02}_3^{\mathsf{T}}, \mathbf{0}_{p_{\text{dep}}-9}^{\mathsf{T}})^{\mathsf{T}}.$$

We considered an administrative censoring with $C_i = 4$, leading to 44% censoring under the three scenarios represented by weak, moderate, and strong correlations of the design matrix.

3.1.3. *Measures of performance* For any $\widehat{\boldsymbol{\beta}} \in \{\widehat{\boldsymbol{\beta}}_{\text{DAC}}, \widehat{\boldsymbol{\beta}}_{\text{full}}, \widehat{\boldsymbol{\beta}}_{\text{full,lin}}, \widehat{\boldsymbol{\beta}}_{\text{MV}}\}$, we report (i) the average computation time for $\widehat{\boldsymbol{\beta}}$; (ii) the global mean squared error (GMSE), defined as $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}} \mathbb{V} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$; (iii) empirical probability of $J \notin \widehat{\mathcal{A}}$; (iv) the bias of each individual coefficient; and (v) mean squared error (MSE) of each individual coefficient. For $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ and $\widehat{\boldsymbol{\beta}}_{\text{full,lin}}$, we also report the empirical coverage level of the 95% normal confidence interval with standard error estimated as described in Section 2.3. For any $\widetilde{\boldsymbol{\beta}} \in \{\widetilde{\boldsymbol{\beta}}_{\text{DAC}}, \widetilde{\boldsymbol{\beta}}_{\Omega_{\text{full}}}\}$, we report (i) the average computation time for $\widetilde{\boldsymbol{\beta}}$; (ii) the global mean squared error (GMSE), defined as $(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}} \mathbb{V} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$.

When only computing time is of interest, we calculate the average computation time for each configuration by averaging over 10 simulated datasets performed on Intel® Xeon® CPU E5-2697 v3 @ 2.60GHz. The average computation time for each configuration in the time-dependent settings is based on simulations using 50 simulated datasets performed on Intel® Xeon® E5-2620 v3 @2.40GHz. The statistical performance is evaluated based on 1000 simulated datasets for each configuration. Although the DAC algorithms can be more efficiently implemented via parallel computing, we report the computational time for DAC algorithms carried out without parallelization for fair comparisons to other algorithms, and for each simulation, we run a single-core job including all the methods under comparison.

3.2. *Simulation results*

We first show in Table 1, the average computation time and GMSE of unpenalized estimators $\widetilde{\boldsymbol{\beta}}_{\text{DAC}}$ and $\widetilde{\boldsymbol{\beta}}_{\Omega_{\text{full}}}$. The results suggest that $\widetilde{\boldsymbol{\beta}}_{\text{DAC}}$ with two iterations ($\mathcal{I} = 2$) attains a GMSE comparable to the full sample-based estimator $\widetilde{\boldsymbol{\beta}}_{\Omega_{\text{full}}}$ and reduced the computation time by more than 50%. The DAC estimator $\widetilde{\boldsymbol{\beta}}_{\text{DAC}}$ with two iterations ($\mathcal{I} = 2$) has a similar GMSE to $\mathcal{I} = 3$. Across all settings, the results of $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ are nearly identical with $\mathcal{I} = 2$ or $\mathcal{I} = 3$, and hence we summarize below the results for $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ only for $\mathcal{I} = 2$ unless noted otherwise.

3.2.1. *Computation time* The computation time is summarized in Tables 2 and 3 for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\text{I})}$ and $\boldsymbol{\beta}_0^{(\text{II})}$, and Tables S1 and S2 of supplementary material available at *Biostatistics* online for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\text{III})}$ for time-independent survival data. There are substantial differences in computation time across methods. Across different settings, the average computation time of $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ ranges from 9.2 to 10.2 s for $p = 50$ and from

19.1 to 58.0 s for $p = 200$, with virtually all time spent on the computation of the unpenalized estimator $\widetilde{\boldsymbol{\beta}}_{\text{DAC}}$. On the contrary, $\widehat{\boldsymbol{\beta}}_{\text{full}}$ requires a substantially longer computation time with average time ranging from 443 to 500 s for $p = 50$ and from 607 to 1080 s for $p = 200$. This suggests that the computation time of $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ is about 2% of the full sample estimator when $p = 50$ and about 5% when $p = 200$. On the other hand, $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ has a substantially longer average computation time than $\widehat{\boldsymbol{\beta}}_{\text{full}}$. This is because the MCP procedure, requiring more computational time than the aLASSO, needs to be fitted $K = 100$ times. As shown in Table S3 of supplementary material available at *Biostatistics* online, when the sample size varies from $n_0 = 10^5$ to $n_0 = 9 \times 10^6$, the computation time increases for all methods but the increase is more drastic for $\widehat{\boldsymbol{\beta}}_{\text{full}}$ and $\widehat{\boldsymbol{\beta}}_{\text{MV}}$. The full sample estimator $\widehat{\boldsymbol{\beta}}_{\text{full}}$ can only be calculated when requesting a very large memory (150GB) when $n_0 = 9 \times 10^6$ while $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ can be computed with a much smaller memory.

In the presence of time-dependent covariates, Table 4 shows that $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ has an average computation time of 112–121 s for $p_{\text{ind}} = 50$ and $p_{\text{dep}} = 50$; $\widehat{\boldsymbol{\beta}}_{\text{full,lin}}$ has an average computation time of 254–264 s. Virtually all computation time for $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ and $\widehat{\boldsymbol{\beta}}_{\text{full,lin}}$ is spent on the computation of the unpenalized initial estimator $\widetilde{\boldsymbol{\beta}}_{\text{DAC}}$, which has more observations and requires substantially more computation time compared to the setting with time-independent covariates given the same n_0 and p.

3.2.2. *Statistical performance* The results for the simulation scenarios with only time-independent covariates are summarized in Tables 2 and 3 for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\text{I})}$ and $\boldsymbol{\beta}_0^{(\text{II})}$, and Tables S1 and S2 of supplementary material available at *Biostatistics* online for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\text{III})}$. In general, $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ is able to achieve a statistical performance comparable to $\widehat{\boldsymbol{\beta}}_{\text{full}}$, while $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ generally has a worse performance, with respect to the GMSE and variable selection, bias, and MSE of individual coefficient. For example, as shown in Table 2, the GMSEs ($\times 10^{-5}$) for $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$, $\widehat{\boldsymbol{\beta}}_{\text{full}}$, and $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ are respectively 4.27, 4.24, and 5.61 when $p = 50$ and $v = 0.2$; 4.1, 4.08, 5.5 when $p = 200$ and $v = 0.2$. The relative performance of different procedures has similar patterns across different levels of correlation v among the covariates. When the signals are relatively strong and sparse as for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\text{I})}$ or $\boldsymbol{\beta}_0^{(\text{II})}$, $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ and $\widehat{\boldsymbol{\beta}}_{\text{full}}$ have small biases and achieved perfect variable selection, while $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ substantially excludes the $\beta_{0j} = 0.05$ signal when $p = 200$. For the more challenging case of $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\text{III})}$ where some of the signals are weak, the variable selection of $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ and $\widehat{\boldsymbol{\beta}}_{\text{full}}$ is also near perfect. Both penalized estimators for the weakest signal (0.035) exhibit a small amount of bias when $v = 0.2$ and 0.5 and an increased bias when $v = 0.8$. Such biases in the weak signals are expected for shrinkage estimators (Menelaos *and others*, 2016), especially in the presence of high correlation among covariates. However, it is important to note that $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ and $\widehat{\boldsymbol{\beta}}_{\text{full}}$ perform nearly identically, suggesting that our DAC_{lin} procedure incurs negligible additional approximation errors. On the other hand, $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ has difficulty in detecting the 0.05 and 0.035 signals and tends to produce substantially higher MSE than $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$.

The empirical coverage levels for the confidence intervals are close to the nominal level across all settings except for the very challenging setting with very weak signals when the correlation is $v = 0.8$. This again is due to the bias inherent in shrinkage estimators. The relative performance of $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$, $\widehat{\boldsymbol{\beta}}_{\text{full}}$ and $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ remains similar across different sample sizes. When n_0 varies from 10^5 to 9×10^6, $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ remains the best performing estimator with precision comparable to $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ while maintaining substantial advantage with respect to computational efficiency.

The results for the time-dependent survival are summarized in Table 4. We find that $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ also generally has a good performance in estimating $\boldsymbol{\beta}_0^{(\text{IV})}$ for both time-independent and time-dependent covariates. The variable selection consistency holds perfectly for all parameters of interest. The coverage of the confidence intervals also has similar patterns as the case with time-independent covariates.

The proposed DAC approach was implemented as an R software package *divideconquer*, which is available at `https://github.com/michaelyanwang/divideconquer`.

Table 1. *Comparisons of* $\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$ $\mathcal{I} = 1, 2, 3$ *and* $\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$ *with respect to average computation time in seconds and global mean squared error (GMSE* $\times 10^{-5}$*) for the estimation of* $\boldsymbol{\beta}_0^{(\mathrm{I})}$, $\boldsymbol{\beta}_0^{(\mathrm{II})}$, $\boldsymbol{\beta}_0^{(\mathrm{III})}$, *and* $\boldsymbol{\beta}_0^{(\mathrm{IV})}$

(a) Time-independent

v	Estimator		$\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\mathrm{I})}$ Time	GMSE	$\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\mathrm{II})}$ Time	GMSE	$\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\mathrm{III})}$ Time	GMSE
					$p = 50$			
0.20	$\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$	$\mathcal{I} = 1$	5.2	19.3	4.5	21.1	4.3	21.2
		$\mathcal{I} = 2$	10.1	19.3	8.7	21.1	8.4	21.2
		$\mathcal{I} = 3$	15.1	19.3	12.9	21.1	12.4	21.2
	$\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$		26.1	19.2	24.0	21.0	24.2	21.1
0.50	$\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$	$\mathcal{I} = 1$	6.3	18.1	5.8	19.5	5.5	20.4
		$\mathcal{I} = 2$	10.3	18.1	9.8	19.5	9.3	20.4
		$\mathcal{I} = 3$	14.3	18.1	13.7	19.5	13.2	20.4
	$\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$		32.9	18.0	31.3	19.4	30.3	20.4
0.80	$\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$	$\mathcal{I} = 1$	5.3	17.9	5.2	18.7	5.1	19.7
		$\mathcal{I} = 2$	9.2	17.8	9.0	18.7	8.9	19.7
		$\mathcal{I} = 3$	13.1	17.8	12.9	18.7	12.8	19.7
	$\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$		30.3	17.7	29.6	18.6	29.2	19.7
					$p = 200$			
0.20	$\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$	$\mathcal{I} = 1$	10.2	74.9	13.9	83.8	17.1	85.0
		$\mathcal{I} = 2$	19.0	74.5	26.4	83.4	32.7	84.6
		$\mathcal{I} = 3$	27.8	74.5	38.8	83.4	48.2	84.6
	$\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$		62.0	74.3	83.1	83.2	102	84.4
0.50	$\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$	$\mathcal{I} = 1$	20.5	69.0	22.8	76.5	24.7	80.6
		$\mathcal{I} = 2$	39.3	68.3	43.9	76.1	47.7	80.3
		$\mathcal{I} = 3$	58.0	68.3	64.8	76.1	70.5	80.3
	$\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$		126	68.1	142	75.9	151	80.1
0.80	$\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$	$\mathcal{I} = 1$	26.9	66.0	28.6	72.0	30.0	77.3
		$\mathcal{I} = 2$	51.9	65.0	55.2	71.5	57.9	76.9
		$\mathcal{I} = 3$	76.8	65.0	81.7	71.5	85.7	76.9
	$\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$		169	64.9	180	71.3	190	76.7

(b) Time-dependent

v	Estimator		$\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\mathrm{IV})}$ *Time*	GMSE
0.2	$\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$	$\mathcal{I} = 1$	62.0	17.9
		$\mathcal{I} = 2$	121	17.9
		$\mathcal{I} = 3$	179	17.9
	$\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$		262	18.0
0.5	$\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$	$\mathcal{I} = 1$	58.0	17.9
		$\mathcal{I} = 2$	112	17.9
		$\mathcal{I} = 3$	166	17.9
	$\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$		263	17.9
0.8	$\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$	$\mathcal{I} = 1$	58.3	18.0
		$\mathcal{I} = 2$	114	18.0
		$\mathcal{I} = 3$	169	18.0
	$\widetilde{\boldsymbol{\beta}}_{\Omega_{\mathsf{full}}}$		254	17.9

Table 2. Comparisons of $\widehat{\boldsymbol\beta}_{\mathrm{DAC}}$ ($\mathcal{I} = 1, 2, 3$), $\widehat{\boldsymbol\beta}_{\mathrm{full}}$, and $\widehat{\boldsymbol\beta}_{\mathrm{MV}}$ for estimating $\boldsymbol\beta_0 = \boldsymbol\beta_0^{(\mathrm{I})}$ with respect to average computation time in seconds, GMSE ($\times 10^{-5}$), coefficient-specific empirical probability (%) of $j \notin \widehat{\mathcal{A}}$, bias ($\times 10^{-4}$), MSE ($\times 10^{-5}$), and empirical coverage probability (%) of the confidence intervals

		$v = 0.2$					$v = 0.5$					$v = 0.8$				
		$\widehat{\boldsymbol\beta}_{\mathrm{DAC}}$			$\widehat{\boldsymbol\beta}_{\mathrm{full}}$	$\widehat{\boldsymbol\beta}_{\mathrm{MV}}$	$\widehat{\boldsymbol\beta}_{\mathrm{DAC}}$			$\widehat{\boldsymbol\beta}_{\mathrm{full}}$	$\widehat{\boldsymbol\beta}_{\mathrm{MV}}$	$\widehat{\boldsymbol\beta}_{\mathrm{DAC}}$			$\widehat{\boldsymbol\beta}_{\mathrm{full}}$	$\widehat{\boldsymbol\beta}_{\mathrm{MV}}$
	$\mathcal{I} =$	1	2	3			1	2	3			1	2	3		
									$p = 50$							
	TIME	5.20	10.2	15.1	443	1451	6.30	10.3	14.3	499	1642	5.30	9.20	13.1	500	1652
	GMSE	4.32	4.27	4.27	4.24	5.61	4.53	4.42	4.42	4.41	7.40	5.17	5.01	5.01	4.97	9.91
$\beta_{0j} = 0.8$	%zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−0.74	0.49	0.49	0.63	12.9	−1.64	−0.24	−0.24	−0.19	12.3	−0.27	1.16	1.16	1.16	14.7
	MSE	0.55	0.55	0.55	0.55	0.72	0.70	0.70	0.70	0.70	0.86	1.50	1.50	1.50	1.50	1.76
	CovP	94.4	94.7	94.7	—	—	94.4	94.7	94.7	—	—	94.6	94.7	94.7	—	—
$\beta_{0j} = 0.4$	%zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−1.47	−0.85	−0.85	−0.72	5.61	−0.63	0.07	0.07	0.03	6.55	−0.02	0.72	0.72	0.73	8.81
	MSE	0.49	0.48	0.48	0.48	0.52	0.55	0.55	0.55	0.55	0.60	1.48	1.48	1.48	1.46	1.60
	CovP	93.0	93.3	93.3	—	—	94.9	95.1	95.1	—	—	95.0	95.1	95.1	—	—
$\beta_{0j} = 0.2$	%zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−1.39	−1.09	−1.09	−1.18	2.75	−0.82	−0.49	−0.49	−0.44	3.88	−3.37	−3.02	−3.02	−2.87	−3.18
	MSE	0.44	0.44	0.44	0.44	0.46	0.63	0.63	0.63	0.63	0.67	1.49	1.48	1.48	1.48	1.76
	CovP	94.5	94.4	94.4	—	—	94.4	94.3	94.3	—	—	94.4	94.4	94.4	—	—
$\beta_{0j} = 0$	%zero	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Bias	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	MSE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
									$p = 200$							
	TIME	10.2	19.1	27.8	607	2246	20.6	39.4	58.1	859	3316	27.0	52.0	76.9	1016	3641
	GMSE	4.62	4.10	4.10	4.08	5.50	5.83	4.61	4.61	4.61	9.50	6.91	5.05	5.05	5.06	17.0
$\beta_{0j} = 0.8$	%zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−5.46	0.31	0.31	0.27	13.2	−6.75	0.25	0.25	0.33	16.0	−6.38	0.94	0.94	0.84	20.7
	MSE	0.58	0.53	0.53	0.53	0.71	0.77	0.70	0.70	0.70	0.97	1.49	1.45	1.45	1.45	1.96
	CovP	94.4	94.7	94.7	—	—	94.2	94.7	94.7	—	—	94.3	94.8	94.8	—	—
$\beta_{0j} = 0.4$	%zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−2.09	0.80	0.80	0.81	7.88	−3.11	0.29	0.29	0.43	8.89	−3.47	0.24	0.24	0.57	11.4
	MSE	0.45	0.44	0.44	0.44	0.51	0.68	0.66	0.66	0.66	0.76	1.42	1.42	1.42	1.42	1.62
	CovP	94.9	95.0	95.0	—	—	94.7	94.8	94.8	—	—	94.8	94.9	94.9	—	—
$\beta_{0j} = 0.2$	%zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−4.15	−2.68	−2.68	−2.69	0.40	−4.15	−2.44	−2.44	−2.52	3.63	−2.02	−0.15	−0.15	−0.09	1.09
	MSE	0.44	0.43	0.43	0.43	0.45	0.61	0.60	0.60	0.60	0.61	1.34	1.33	1.33	1.33	1.62
	CovP	94.5	94.6	94.6	—	—	94.5	94.6	94.6	—	—	94.7	94.9	94.9	—	—
$\beta_{0j} = 0$	%zero	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Bias	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	MSE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3. Comparisons of $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$ ($\mathcal{I} = 1, 2, 3$), $\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$, and $\widehat{\boldsymbol{\beta}}_{\mathsf{MV}}$ for estimating $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\mathrm{II})}$ with respect to average computation time in seconds, GMSE ($\times 10^{-5}$), coefficient-specific empirical probability (%) of $j \notin \widehat{\mathcal{A}}$, bias ($\times 10^{-4}$), MSE ($\times 10^{-5}$), and empirical coverage probability (%) of the confidence intervals

		$v = 0.2$					$v = 0.5$					$v = 0.8$				
		$\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$			$\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$	$\widehat{\boldsymbol{\beta}}_{\mathsf{MV}}$	$\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$			$\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$	$\widehat{\boldsymbol{\beta}}_{\mathsf{MV}}$	$\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$			$\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$	$\widehat{\boldsymbol{\beta}}_{\mathsf{MV}}$
	$\mathcal{I} =$	1	2	3			1	2	3			1	2	3		
								$p = 50$								
	TIME	4.5	8.7	12.9	461	1554	5.9	9.8	13.7	497	1628	5.2	9.1	12.9	499	1657
	GMSE	7.27	7.24	7.24	7.22	870	6.92	6.86	6.86	6.84	1176	7.21	7.12	7.12	7.17	4452
$\beta_{0j} = 0.4$	%zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	1.14	1.59	1.59	1.60	97.8	1.47	2.07	2.07	2.11	110	4.82	5.45	5.45	1.39	219
	MSE	0.53	0.53	0.53	0.53	10.1	0.71	0.71	0.71	0.71	12.8	1.70	1.70	1.70	1.68	49.6
	CovP	94.4	94.8	94.8	—	—	94.5	94.9	94.9	—	—	94.5	94.8	94.8	—	—
$\beta_{0j} = 0.2$	%zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	0.72	0.94	0.94	1.14	98.3	1.61	1.90	1.90	2.02	108	3.63	3.95	3.95	1.06	206
	MSE	0.50	0.50	0.50	0.50	10.2	0.76	0.76	0.76	0.75	12.4	1.65	1.65	1.65	1.66	44.2
	CovP	95.0	95.1	95.1	—	—	95.0	95.1	95.1	—	—	95.0	95.1	95.1	—	—
$\beta_{0j} = 0.05$	%zero	0	0	0	0	61.6	0	0	0	0	100	0	0	0	0	100
	Bias	−4.21	−4.16	−4.16	−4.21	−407	−5.65	−5.55	−5.55	−5.52	−500	−11.4	−11.3	−11.3	−7.22	−500
	MSE	0.50	0.50	0.50	0.49	180	0.71	0.71	0.71	0.70	250	2.08	2.07	2.07	2.05	250
	CovP	95.2	95.2	95.2	—	—	94.8	94.9	94.9	—	—	91.5	91.7	91.7	—	—
$\beta_{0j} = 0$	%zero	100	100	100	100	100	100	100	100	100	100	99.9	100	100	100	100
	Bias	0	0	0	0	0	0	0	0	0	0	0.09	0	0	0	0
	MSE	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0
								$p = 200$								
	TIME	14.0	26.5	38.9	696	2755	22.9	43.9	64.9	920	3465	28.7	55.3	81.8	1051	3764
	GMSE	7.53	7.29	7.29	7.28	1496	7.58	7.05	7.05	7.03	1300	8.38	7.46	7.46	7.67	12696
$\beta_{0j} = 0.4$	%zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−0.02	2.00	2.00	2.07	139	0.84	3.59	3.59	3.58	125.2	5.70	8.81	8.81	0.58	172
	MSE	0.54	0.54	0.54	0.54	19.8	0.80	0.79	0.79	0.79	16.5	1.62	1.67	1.67	1.62	31.2
	CovP	95.4	94.9	94.9	—	—	94.8	94.9	94.9	—	—	96.3	96.2	96.2	—	—
$\beta_{0j} = 0.2$	%zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−0.02	0.99	0.99	1.08	142	0.15	1.59	1.59	1.77	122	3.94	5.37	5.37	5.56	151
	MSE	0.49	0.49	0.49	0.49	20.8	0.75	0.75	0.75	0.74	15.8	1.72	1.73	1.73	1.79	25.1
	CovP	96.4	96.4	96.4	—	—	95.4	95.2	95.2	—	—	95.1	95.3	95.3	—	—
$\beta_{0j} = 0.05$	%zero	0	0	0	0	99.1	0	0	0	0	100	0	0	0	0	100
	Bias	−6.39	−6.07	−6.07	−6.08	−498	−8.32	−7.92	−7.92	−7.83	−500	−18.5	−18.1	−18.1	−15.4	−500
	MSE	0.59	0.57	0.57	0.57	249	0.84	0.82	0.82	0.82	250	2.38	2.38	2.38	2.43	250
	CovP	93.6	93.2	93.2	—	—	93.1	93.5	93.5	—	—	90.5	91.0	91.0	—	—
$\beta_{0j} = 0$	%zero	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Bias	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	MSE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4. Performance of $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$ ($\mathcal{I} = 1, 2, 3$) and $\widehat{\boldsymbol{\beta}}_{\mathsf{full,lin}}$ for estimating $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{(\mathsf{IV})}$ with respect to average computation time in seconds, GMSE ($\times 10^{-5}$), coefficient-specific empirical probability (%) of $j \notin \widehat{\mathcal{A}}$, bias ($\times 10^{-4}$), MSE ($\times 10^{-5}$), and empirical coverage probability (%) of the confidence intervals

		$\nu = 0.2$				$\nu = 0.5$				$\nu = 0.8$			
		$\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$			$\widehat{\boldsymbol{\beta}}_{\mathsf{full,lin}}$	$\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$			$\widehat{\boldsymbol{\beta}}_{\mathsf{full,lin}}$	$\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$			$\widehat{\boldsymbol{\beta}}_{\mathsf{full,lin}}$
	$\mathcal{I} =$	1	2	3		1	2	3		1	2	3	
	Time	62.0	121	179	262	58.0	112	166	264	58.3	114	169	254
	GMSE	3.55	3.55	3.55	3.55	3.68	3.68	3.68	3.69	4.44	4.44	4.44	4.44
	Time-independent covariates												
$\beta_{0j} = 0.08$	%zero	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	1.45	1.46	1.46	1.46	4.62	4.63	4.63	4.65	11.9	11.9	11.9	12.0
	MSE	0.22	0.22	0.22	0.22	0.36	0.36	0.36	0.36	1.02	1.02	1.02	1.03
	CovP	95.7	95.7	95.7	95.8	94.6	94.6	94.6	94.7	93.9	93.9	93.9	93.5
$\beta_{0j} = 0.04$	%zero	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−0.19	−0.19	−0.19	−0.14	0.53	0.54	0.54	0.61	3.45	3.47	3.47	3.45
	MSE	0.22	0.22	0.22	0.22	0.33	0.33	0.33	0.33	0.93	0.93	0.93	0.93
	CovP	95.4	95.5	95.5	95.5	96.0	96.1	96.1	96.0	93.7	93.8	93.8	93.7
$\beta_{0j} = 0.02$	%zero	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−4.55	−4.56	−4.56	−4.58	−6.79	−6.78	−6.78	−6.89	−17.4	−17.4	−17.4	−17.2
	MSE	0.25	0.25	0.25	0.25	0.45	0.45	0.45	0.45	1.59	1.59	1.59	1.57
	CovP	94.1	94.1	94.1	94.4	90.5	90.8	90.8	90.8	86.8	86.7	86.7	86.8
$\beta_{0j} = 0$	%zero	100	100	100	100	100	100	100	100	100	100	100	100
	Bias	0	0	0	0	0	0	0	0	0	0	0	0
	MSE	0	0	0	0	0	0	0	0	0	0	0	0
	Time-independent covariates												
$\beta_{0j} = 0.08$	%zero	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	1.01	1.01	1.01	1.04	2.65	2.65	2.65	2.61	9.72	9.74	9.74	9.74
	MSE	0.21	0.21	0.21	0.21	0.37	0.37	0.37	0.38	0.93	0.93	0.93	0.95
	CovP	96.3	96.3	96.3	96.3	94.6	94.6	94.6	94.4	95.6	95.5	95.5	95.4
$\beta_{0j} = 0.04$	%zero	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−1.01	−1.01	−1.01	−1.01	1.11	1.11	1.11	1.15	3.26	3.26	3.26	3.19
	MSE	0.21	0.21	0.21	0.21	0.33	0.33	0.33	0.33	0.93	0.93	0.93	0.94
	CovP	95.8	95.9	95.9	96.0	95.2	95.2	95.2	95.3	94.9	94.8	94.8	95.0
$\beta_{0j} = 0.02$	%zero	0	0	0	0	0	0	0	0	0	0	0	0
	Bias	−3.81	−3.81	−3.81	−3.81	−6.28	−6.26	−6.26	−6.30	−16.5	−16.5	−16.5	−16.6
	MSE	0.26	0.26	0.26	0.25	0.45	0.45	0.45	0.45	1.49	1.48	1.48	1.49
	CovP	93.8	93.8	93.8	93.5	91.2	91.1	91.1	91.4	86.2	86.1	86.1	86.0
$\beta_{0j} = 0$	%zero	100	100	100	100	100	100	100	100	100	100	100	100
	Bias	0	0	0	0	0	0	0	0	0	0	0	0
	MSE	0	0	0	0	0	0	0	0	0	0	0	0

4. Application of the DAC procedure to Medicare data

We applied the proposed DAC_{lin} algorithm to develop risk prediction models for heart failure-specific readmission or death within 30 days of discharge among Medicare patients who were admitted due to heart failure. The Medicare inpatient claims were assembled for all Medicare fee-for-service beneficiaries during $2000-2012$ to identify the eligible study population. The index date was defined as the discharge date of the first heart failure admission of each patient. We restricted the study population to patients who were discharged alive from the first heart failure admission. The outcome of interest was time to heart failure-specific readmission or death after the first heart failure admission. Because readmission rates within 30 days were used to assess the quality of care at hospitals by the Centers for Medicare and Medicaid Services (CMS) (CMS, 2016), we censored the time to readmission at 30 days. For a patient who was readmitted or died on the same day as discharge (whose claim did not indicate discharge dead), the time-to-event was set at 0.5 days. Due to the large number of ICD-9 codes, we classified each discharge ICD-9 code into disease phenotypes indexed by phenotype codes according to Denny *and others* (2013). A heart failure admission or readmission was identified if the claim for that admission or readmission had a heart failure phenotype code at discharge.

We considered two sets of covariates: (i) time-independent covariates including baseline individual-level covariates collected at time of discharge from the index heart failure hospitalization, baseline area-level covariates at the residential ZIP code of each patient, and indicators for time trend including dummy variables for each year and dummy variables for each months, and (ii) time-dependent predictors that vary day-by-day. Baseline individual-level covariates included age, sex, race (white, black, others), calendar year, and month of the discharge, Charlson comorbidity index (CCI) (Quan *and others*, 2005) which described the degree of illness of a patient, and indicators for nonrare comorbidities (defined as prevalence > 0.1 among the study population). Baseline area-level covariates included socioeconomic status variables (percent black residents [ranging from 0 to 1], percent Hispanic residents [ranging from 0 to 1], median household income [per 10 000 increase], median home value [per 10 000 increase], percent below poverty [ranging from 0 to 1], percent below high school [ranging from 0 to 1], percent owned houses [ranging from 0 to 1]), population density (1000 per squared kilometer), and health status variables (percent taking hemoglobin A1C test [ranging from 0 to 1], average BMI, percent ambulance use [ranging from 0 to 1], percent having low-density lipoprotein test [ranging from 0 to 1], and smoke rate [ranging from 0 to 1]). The time-dependent covariates included daily fine particulate matter ($\text{PM}_{2.5}$) predicted using a neural network algorithm (Di *and others*, 2016), daily temperature with its quadratic form, and daily dew point temperature with its quadratic form. There were 574 time-independent covariates and five time-dependent covariates.

There were $n_0 = 9\,567\,752$ eligible patients with a total of $d_0 = 2\,079\,436$ heart failure readmissions or deaths, among which 1 453 627 were readmissions and 625 809 were deaths. After expanding the dataset by accounting for time-dependent variables which vary day-by-day, the time-dependent dataset contained 245 623 834 rows of records.

We fitted cause-specific Cox models for readmission due to heart failure or deaths as a composite outcome, considering two separate models: (i) a model containing only time-independent covariates and (ii) a model incorporating time-dependent covariates. In both cases, the datasets were too large for `glmnet` package to analyze as a whole, demonstrating the need for DAC_{lin}.

4.1. *Time-independent covariates only*

We applied DAC_{lin} with $K = 50$ and paralleled DAC_{lin} on 25 Authentic AMD Little Endian @2.30GHz CPUs. Computing $\widehat{\boldsymbol{\beta}}_{\text{DAC}}$ with $\mathcal{I} = 2$ took 1.1 h, including the time of reading datasets from hard drives during each iteration of the update of the one-step estimator. Figure 1a shows the hazard ratio of each

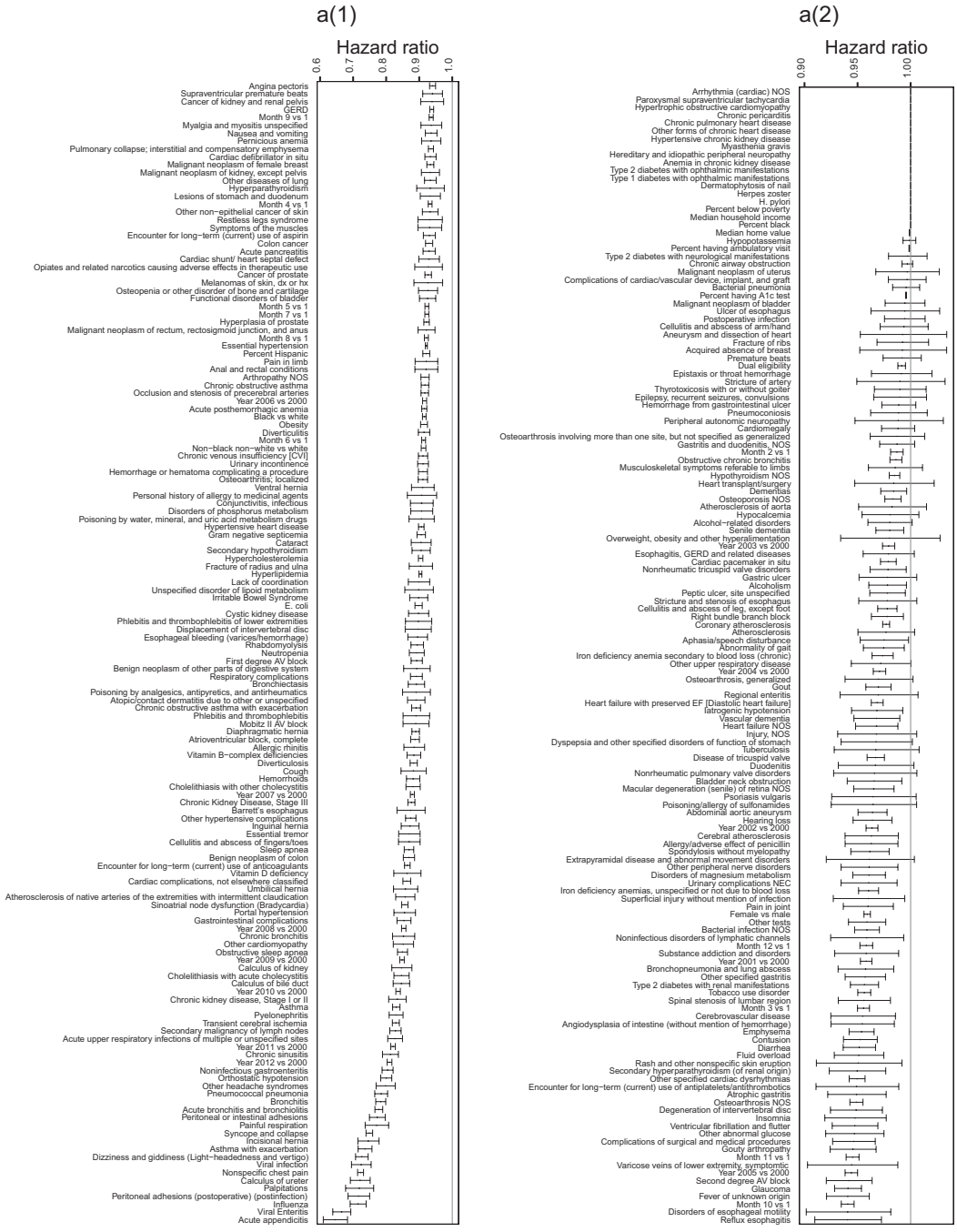

Fig. 1. Hazard ratios of each covariate in estimating hazard of heart failure readmissions or death within 30 days after the first admission using DAC_{lin}. (a) time-independent variables, (a) time-independent variables (continue), (b) time-dependent variables, and (b) time-dependent variables (continue).

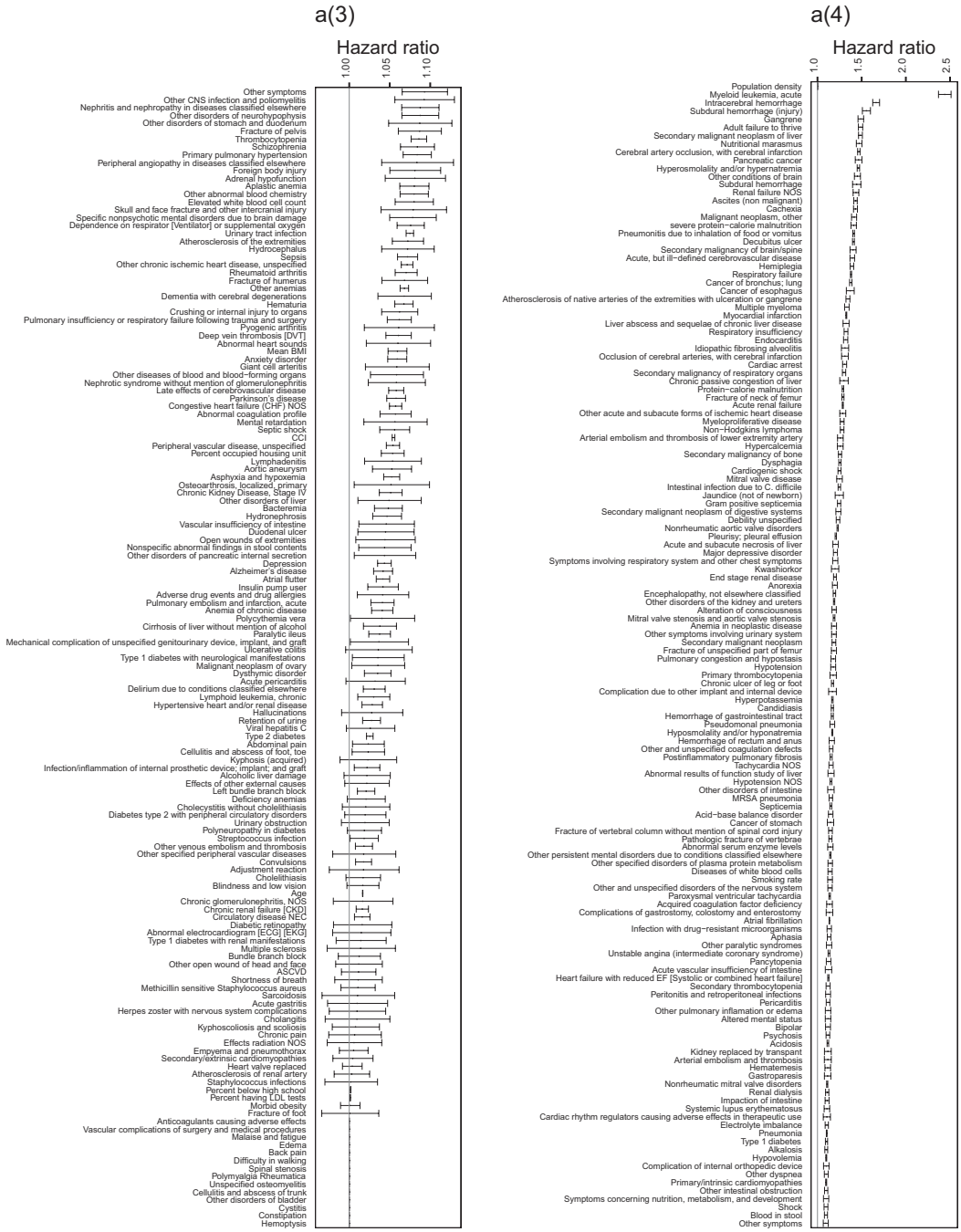

Fig. 1. Continued. a(3) and a(4) time-independent variables.

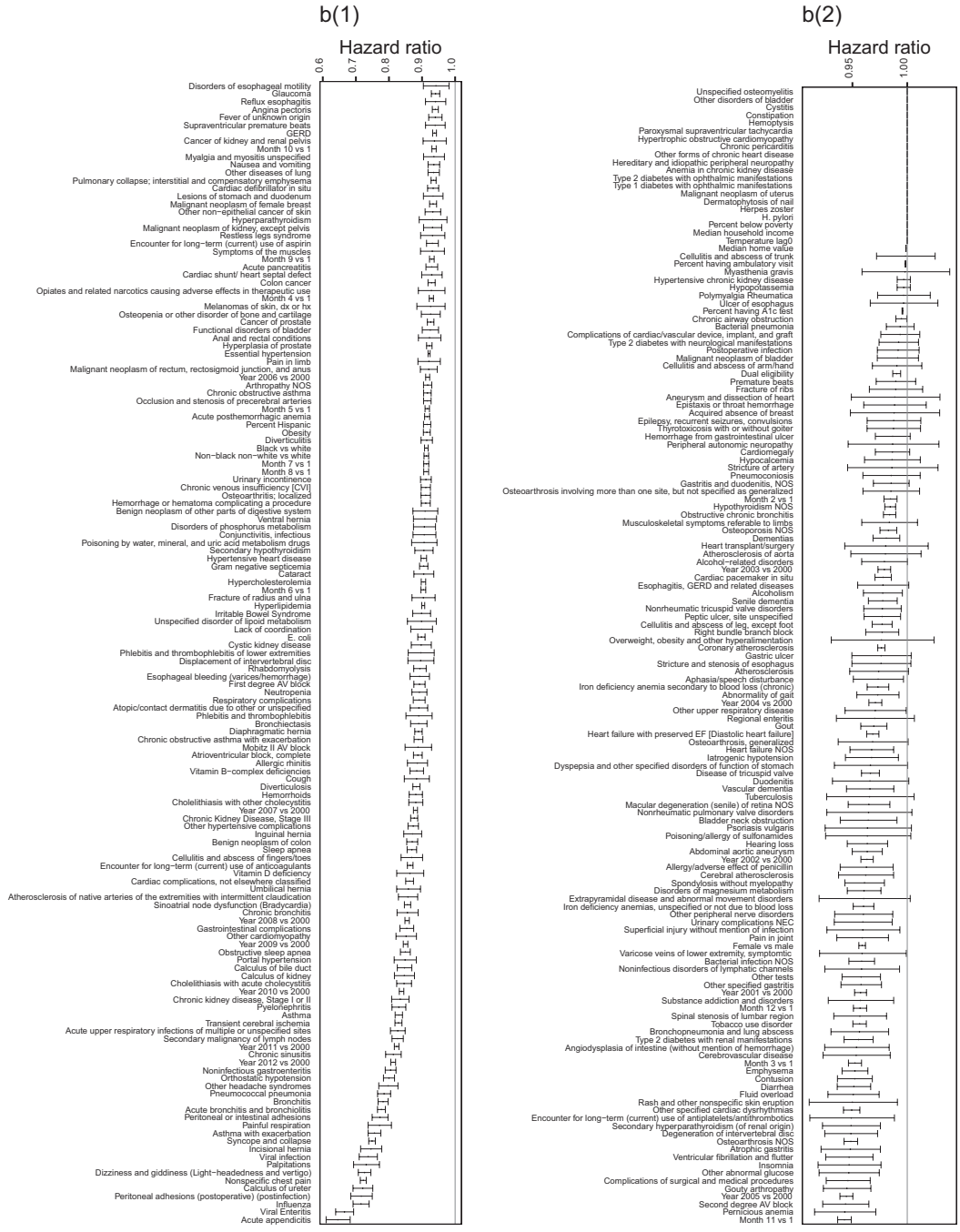

Fig. 1. Continued. b(1) and b(2) time-dependent variables.

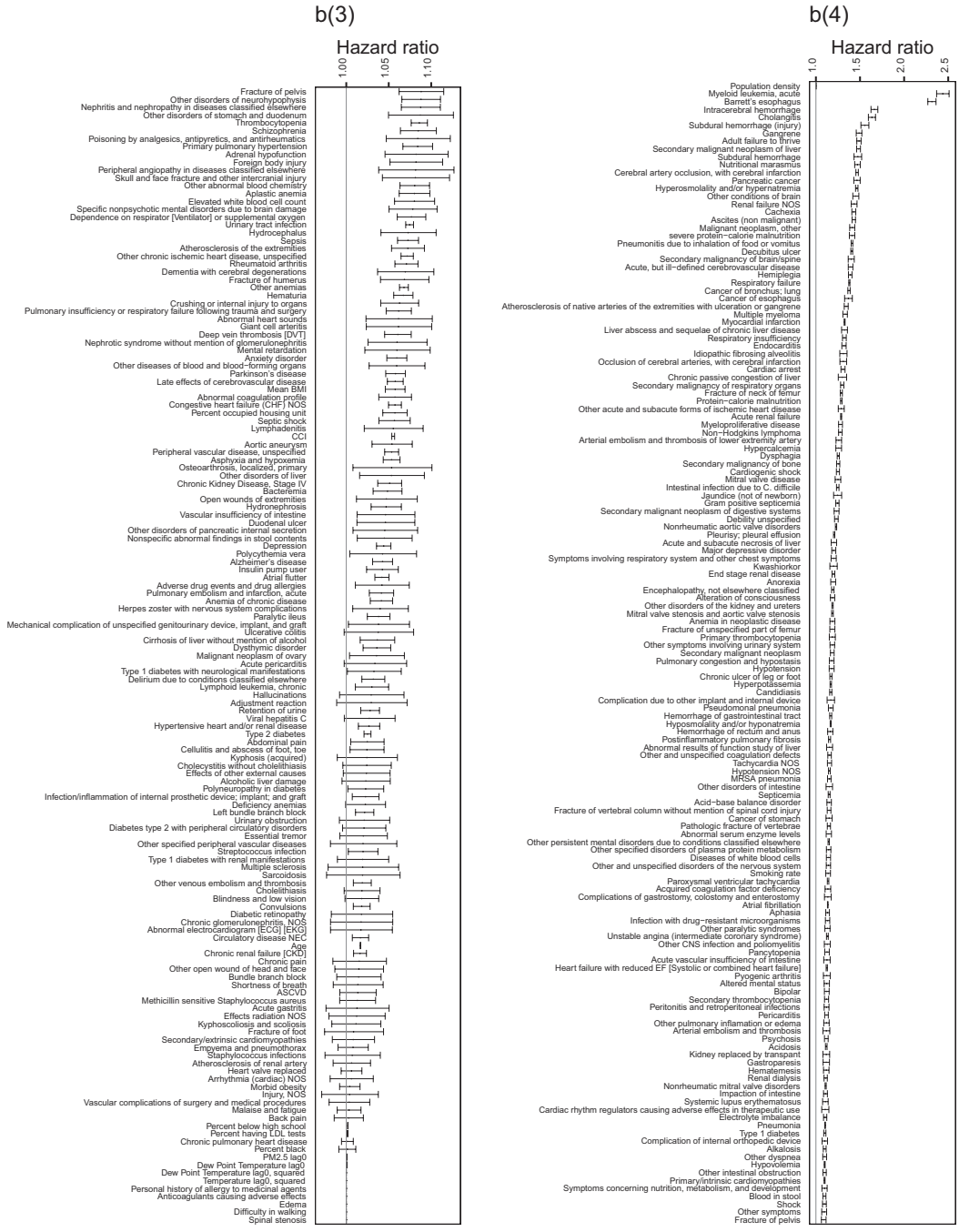

Fig. 1. Continued. b(3) and b(4) time-dependent variables.

covariate based on $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$ with $\mathcal{I} = 2$ predicting heart failure-specific readmission and death within 30 days.

Multiple comorbidities are associated with an increased risk of 30-day readmission or death with the leading factors including renal failures, cancers, malnutrition, subdural or intracerebral hemorrhage, myocardial infarction, endocarditis, respiratory failure, and cardiac arrest. CCI is also associated with an increased hazard of the outcome. These findings are generally consistent with those reported in the literature. For example, Philbin and DiSalvo (1999) reported that ischemic heart disease, diabetes, renal diseases, and idiopathic cardiomyopathy were associated with an increased risk of heart failure-specific readmission within a year. Leading factors negatively associated with readmissions include virus infections, asthma, and chronic kidney disease in earlier stages. These negative association findings are reflective of both clinical practice patterns and the biological effects, as most of the negative predictors are generally less severe than the positive predictors.

Some socioeconomic status predictors are relatively less important in predicting the outcome after accounting for the phenotypes, where percentage of black people, median household income, and percent below poverty are dropped and dual eligibility, median home value, percent below high school has a small hazard ratio. By comparison, Philbin *and others* (2001) reported a decrease in readmission as neighborhood income increased. Foraker *and others* (2011) reported that given comorbidity measured by CCI, the readmission or death hazard was higher for low socioeconomic status patients. The present article considered more detailed phenotypes in addition to CCI suggesting a relatively smaller impact of socioeconomic status. The difference in results is possibly because comorbidity may be on the causal pathway between socioeconomic status and readmission or death. Adjusting for a detailed set of comorbidities partially blocks the effect of socioeconomic status. Percent Hispanic residents is negatively associated with readmission or death. Percent occupied houses increase the risk of readmission or death, which is consistent with the strong positive prediction by population density. Most ecological health variables showed a small hazard ratio.

Black and other race groups have a lower hazard than white. Females have a lower hazard than males, which is consistent with Roger *and others* (2004) that females had a higher survival rate than males after heart failure. Age is associated with an increased hazard of readmission or death, as expected.

The coefficient by month suggests a higher risk of readmission or death in cold seasons than warm seasons, with a larger negative hazard ratio for summer indicators. The short-term readmission or death rate is decreasing over time, which is suggested by the negative hazard ratio of later years. The later calendar year being negatively associated with readmission risk may be an indication of improved follow-up care for patients discharged from heart failure. Consistently, (Roger *and others*, 2004) also suggested an improved heart failure survival rate over time.

4.2. *Incorporating time-dependent covariates*

The analysis in Section 4.2 has two goals. First, the covariates serve as the risk predictors of the hazard of heart failure-specific readmission. Second, all covariates other than $PM_{2.5}$ serve as the potential confounders of the association between $PM_{2.5}$ and readmission, particularly time trend and area-level covariates. The DAC_{lin} procedure is a variable selection technique to drop non-informative confounders given the high dimensionality of confounders. This goal aligns with Belloni *and others* (2014), which constructs separate penalized regressions for the propensity score model and outcome regression model to identify confounders. We, herein, focused on building a penalized regression for the outcome regression model.

We applied DAC_{lin} algorithm with $K = 200$ to this time-dependent survival dataset. The procedure was paralleled on 10 Authentic AMD Little Endian @2.30GHz CPUs. The estimation of $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$ with $\mathcal{I} = 2$ took 36.5 h, including the time of loading the datasets into memory. The result suggests each 10 $\mu g\ m^{-3}$

increase in daily $PM_{2.5}$ was associated with 0.5% increase of risk (95% confidence interval: 0.3–0.7) adjusting for individual-level, area-level covariates, and temperature. Because there is rare evidence on whether air pollution is associated with heart failure-specific readmission or death among heart failure patients and it is rare to estimate the health effect of daily air pollution using a time-dependent Cox model, this model provides a novel approach to address a new research question. While evidence is rare on the association between daily $PM_{2.5}$ and heart failure-specific readmission, some studies used case-crossover design to estimate the effect of short-term $PM_{2.5}$ on the incidence of heart failure admissions. Pope *and others* (2008) found that a 10 $\mu g\ m^{-3}$ increase in 14-day moving average $PM_{2.5}$ was associated with a 13.1 (1.3–26.2) increase in the incidence of heart failure admissions among elderly patients; Zanobetti *and others* (2009) reported that each 10 $\mu g\ m^{-3}$ increase in 2-day averaged $PM_{2.5}$ was associated with a 1.85 (1.19–2.51) increase in the incidence of congestive heart failure admission. There is also a large body of literature suggesting that short-term exposure to $PM_{2.5}$ is associated with an increased risk of death. For example, Di *and others* (2017) shows among the Medicare population during 2000–2012 that each 10 $\mu g\ m^{-3}$ increase in $PM_{2.5}$ was associated with an 1.05% (0.95–1.15) increase in mortality risk. In addition, Figure 1b shows the covariate-specific estimates of the hazard ratio for all the covariates, with the estimates consistent with the analysis of time-independent dataset.

5. Discussion

The proposed DAC_{lin} procedure for fitting aLASSO penalized Cox model reduces the computation cost, while it maintains the precision of estimation and accuracy in variable selection with an extraordinarily large n_0 and a numerically large p. The use of $\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}$ makes it feasible to obtain the $\sqrt{n_0}$-consistent estimator required by the penalized step (e.g. when there is a constraint in RAM) and shortens the computation time of the initial estimator by $>$ 50%. The improvement in the computation time was substantial in the regularized regression step. The LSA converted the fitting of regularized regression from using a dataset of size n_0 to a dataset of size p.

The majority voting-based method $\widehat{\boldsymbol{\beta}}_{\mathsf{MV}}$ with MCP (Chen and Xie, 2014) had a substantially longer computation time than $\widehat{\boldsymbol{\beta}}_{\mathsf{full}}$. The difference primarily comes from (i) the fact that the Cox model with MCP is fitted K times and (ii) the computational efficiency between `glmnet` algorithm which is more efficient than the MCP algorithm in `ncvsurv` (Breheny and Huang, 2011).

The difference in variable selection between $\widehat{\boldsymbol{\beta}}_{\mathsf{DAC}}$ and $\widehat{\boldsymbol{\beta}}_{\mathsf{MV}}$ (Chen and Xie, 2014) is primarily due to the majority voting. The simulations in Chen and Xie (2014) have shown that an increase in the percentage for the majority voting decreased the sensitivity and increased the specificity of variable selection. Similarly in the simulations of the present study with a 50% of the majority vote, Chen and Xie's procedure showed a high specificity but the sensitivity was low for weaker signals as demonstrated in the simulation studies.

For non-weak signals, the oracle property appears to hold well as evidenced by the simulation results for $\boldsymbol{\beta}_0^{(\mathrm{I})}$ and $\boldsymbol{\beta}_0^{(\mathrm{II})}$ shown in Tables 2 and 3. For weak signals such as 0.035 in $\boldsymbol{\beta}_0^{(\mathrm{III})}$, the oracle property does not appear to hold even with $n_0 = 1\,000\,000$ and the bias in the shrinkage estimators results in confidence intervals with low coverage. This is consistent with the impossibility result shown in Potscher and Schneider (2009), which suggests difficulty in deriving precise interval estimators when aLASSO penalty is employed.

When parallel computing is available, a larger K may be preferable for our algorithm as it can reduce the overall computational time provided that $K = o(n_0^{\frac{1}{2}})$. When n is not large relative to p, we may increase the stability of the algorithm by replacing $\widetilde{\boldsymbol{\beta}}_{\mathsf{DAC}}^{[0]}$ with $\widetilde{\boldsymbol{\beta}}_{\Omega_1 \cup \cdots \Omega_d}$ for some d such that $nd \gg p$. Potential approaches to further reduce the computation burden for large p settings include employing a screening step or employing DAC in the second step of the algorithm. However, deriving the statistical properties of DAC algorithms in the large p setting can be more involved and warrants further research.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

USEPA (RD-83587201), an National Institutes of Health (U54 HG007963), and National Institute of Environmental Health Sciences (R01 ES024332 and P30 ES000002). Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication. The analysis of the Medicare data was performed on the secured clusters of Research Computing, Faculty of Arts and Sciences at Harvard University.

REFERENCES

AUSTIN, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine* **31**, 3946–3958.

BELLONI, A., CHERNOZHUKOV, V. AND HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81**, 608–650.

BREHENY, P. AND HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* **5**, 232–253.

CHEN, X. AND XIE, M.-G. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* **24**, 1655–1684.

CMS. (2016). *Readmissions Reduction Program* https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html (accessed November 2, 2017).

COX, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B (Methodological)* **34**, 87–22.

DENNY, J. C., BASTARACHE, L., RITCHIE, M. D., CARROLL, R. J., ZINK, R., MOSLEY, J. D., FIELD, J. R., PULLEY, J. M., RAMIREZ, A. H., BOWTON, E. *and others*. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology* **31**, 1102–1110.

DI, Q., DAI, L., WANG, Y., ZANOBETTI, A., CHOIRAT, C., SCHWARTZ, J. D., DOMINICI, F. *and others* (2017). Association of short-term exposure to air pollution with mortality in older adults. *Journal of the American Medical Association* **318**, 2446–2456.

DI, Q., KLOOG, I., KOUTRAKIS, P., LYAPUSTIN, A., WANG, Y. AND SCHWARTZ, J. (2016). Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environmental Science Technology* **50**, 4712–4721.

EFRON, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72**, 557–565.

FORAKER, R. E., ROSE, K. M., SUCHINDRAN, C. M., CHANG, P. P., MCNEILL, A. M. AND ROSAMOND, W. D. (2011). Socioeconomic status, Medicaid coverage, clinical comorbidity, and rehospitalization or death after an incident heart failure hospitalization. *Circulation Heart Failure* **4**, 308.

FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1.

GOEMAN, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* **52**, 70–84.

PAVLOU, M., AMBLER, G., SEAMAN, S., DE IORIO, M. AND OMAR, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine* **35**, 1159–1177.

PARK, M. Y. AND HASTIE, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 659–677.

PHILBIN, E. F., DEC, G. W., JENKINS, P. L. AND DISALVO, T. G. (2001). Socioeconomic status as an independent risk factor for hospital readmission for heart failure. *The American Journal of Cardiology* **87**, 1367–1371.

PHILBIN, E. F. AND DISALVO, T. G. (1999). Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology* **33**, 1560–1566.

POPE, C. A., RENLUND, D. G., KFOURY, A. G., MAY, H. T. AND HORNE, B. D. (2008). Relation of heart failure hospitalization to exposure to fine particulate air pollution. *The American Journal of Cardiology* **102**, 1230–1234.

POTSCHER, B. M. AND SCHNEIDER, U. (2009). On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference* **139**, 2775–2790.

QUAN, H., SUNDARARAJAN, V., HALFON, P., FONG, A., BURNAND, B., LUTHI, J.-C., SAUNDERS, L. D., BECK, C. A., FEASBY, T. E. AND GHALI, W. A. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care* **43**, 1130–1139.

R CORE TEAM. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

ROGER, V. L., WESTON, S. A., REDFIELD, M. M., HELLERMANN-HOMAN, J. P., KILLIAN, J., YAWN, B. P., JACOBSEN, S. J. *and others* (2004). Trends in heart failure incidence and survival in a community-based population. *Journal of the American Medical Association* **292**, 344–350.

SIMON, N., FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* **39**, 1–13.

TANG, L., ZHOU, L. AND SONG, P. X.-K. (2016). Method of divide-and-combine in regularised generalised linear models for big data. *arXiv preprint arXiv:1611.06208*.

VOLINSKY, C. T. AND RAFTERY, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56**, 256–262.

WANG, H. AND LENG, C. (2007). Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association* **102**, 1039–1048.

ZANOBETTI, A., FRANKLIN, M., KOUTRAKIS, P. AND SCHWARTZ, J. (2009). Fine particulate air pollution and its components in association with cause-specific emergency admissions. *Environmental Health* **8**, 58.

ZHANG, H. H. AND LU, W. (2007). Adaptive LASSO for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

[*Received March 27, 2018; revised August 25, 2019; accepted for publication August 25, 2019*]