Bayesian Cox Regression for Population-scale Inference in Electronic Health Records

Alexander W. Jung\textsuperscript{1,2} and Moritz Gerstung\textsuperscript{1,3,4}

\textsuperscript{1}European Bioinformatics Institute, EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK
\textsuperscript{2}University of Cambridge, Trumpington St, Cambridge CB2 1RL, UK
\textsuperscript{3}Genome Biology Unit, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany
\textsuperscript{4}German Cancer Research Center, dkfz, Neuenheimer Feld 280, 69120 Heidelberg, Germany

June 21, 2021

Abstract

The Cox model is an indispensable tool for time-to-event analysis, particularly in biomedical research. However, medicine is undergoing a profound transformation, generating data at an unprecedented scale, which opens new frontiers to study and understand diseases. With the wealth of data collected, new challenges for statistical inference arise, as datasets are often high dimensional, exhibit an increasing number of measurements at irregularly spaced time points, and are simply too large to fit in memory. Many current implementations for time-to-event analysis are ill-suited for these problems as inference is computationally demanding and requires access to the full data at once. Here we propose a Bayesian version for the counting process representation of Cox’s partial likelihood for efficient inference on large-scale datasets with millions of data points and thousands of time-dependent covariates. Through the combination of stochastic variational inference and a reweighting of the log-likelihood, we obtain an approximation for the posterior distribution that factorizes over subsamples of the data, enabling the analysis in big data settings. Crucially, the method produces viable uncertainty estimates for large-scale and high-dimensional datasets. We show the utility of our method through a simulation study and an application to myocardial infarction in the UK Biobank.

Keywords: Survival Analysis, Time-dependent Covariates, High-dimensional data, Batch learning, Variational Inference
1 Introduction

A fundamental aspect of epidemiology is to identify and study the factors driving disease. One early example is the Framingham Heart Study, which shaped our understanding of cardiovascular outcomes and established risk factors like blood pressure and cholesterol that have become routine in standard health assessments (Dawber et al. 1951).

The possibilities and availability for this type of analysis have dramatically increased with the profound transformation of the biomedical sector in the last decades through digitization and the unprecedented scale of data generation. Electronic health records (EHR), containing the medical history of individuals systematically collected for millions of people and sometimes whole populations, provide new opportunities on how we can approach diseases.

Some recent studies utilizing large-scale EHR with around 1-20 million individuals include: Clift et al. (2020) and Williamson et al. (2020) studying the risk factors associated with severe outcomes for Covid19, Hippisley-Cox et al. (2017) predicting risk of cardiovascular disease or Hippisley-Cox & Coupland (2021) estimating the risk of prostate cancer in asymptomatic men.

The statistical method underlying these studies is the Cox partial likelihood, introduced by Cox (1972, 1975), as the main objective of interest is the time passed until the occurrence of an event.

While estimation with a simple Cox model is possible in large-scale settings, as the aforementioned examples show, it becomes infeasible for more complex analyses often encountered in EHR e.g. high-dimensional data, time-varying covariates, and multiple events. In many cases, one is interested in the evolving status of patients, particularly when multiple measurements over time are available, highlighting the need for a version of the Cox model that scales to the vast amounts of data while simultaneously being able to handle many of the statistical challenges arising in EHR data analysis.
1.1 Related Work

A comprehensive treatment of Bayesian methods for time-to-event analysis can be found in Ibrahim et al. (2001), with applications in Alvares et al. (2021). Most implementations for the Cox model explicitly estimate the baseline hazard through a non-parametric prior, with choices including the Gamma process (Kalbfleisch 1978, Sinha 1993), Beta Process, (Hjort 1990, Laud et al. 1998), or correlated prior processes (Qiou et al. 1999), but also spline functions (Sharef et al. 2010). While specifying the baseline hazard has advantages, like absolute risk prediction, the traditional Cox model circumvents the explicit estimation. Hjort (1990) and Sinha (2003) provide justifications for a Bayesian version of the Cox model without the need to estimate the baseline hazard, closely resembling the traditional Cox model. Recent Bayesian methods for variable selection in high-dimensional data include Shin et al. (2018) and Nikooienejad et al. (2020), however, they have mostly focused on fixed covariates. Generally, none of the Bayesian methods have focused on large-scale applications.

From a frequentist point of view, in the context of high-dimensional data, the Lasso or \(\ell_1\) regularization proposed by Tibshirani (1996, 1997) has been a popular choice. An overview of time-to-event analysis with high-dimensional data more generally can be found in Witten & Tibshirani (2010). Computationally efficient implementations for the Cox model under penalization have been developed by Friedman et al. (2010), Simon et al. (2011), Mittal et al. (2014) and Yang & Zou (2013). Similar to the Bayesian versions, these methods have mainly focused on fixed covariates, while the R software package \texttt{R-glmnet} by Friedman et al. (2010) and Simon et al. (2011) has recently added support for a Cox model based on the counting process representation in combination with the Lasso, enabling the inclusion of time-varying covariates. However, the proposed methods require the full data for optimization, limiting the application to moderately sized datasets.

More recently, approaches that factorize the likelihood over subsamples of the data
have been proposed, facilitating time-to-event analysis at large scales. Li et al. (2020) proposes a solution based on the batch screening iterative lasso algorithm as an extension to the traditional Lasso. Kvamme et al. (2019) suggest a modified version of the Cox model based on a hypothetical weighting of the likelihood to justify training on subsamples via stochastic gradient descent (SGD). Tarkhan & Simon (2020) provide a more general treatment of the combination of Cox’s partial likelihood with SGD.

While these methods can handle large-scale datasets, they cannot incorporate time-varying covariates measured at different time points, nor do they provide standard errors or other measures of uncertainty. The last method does propose a plug-in estimator or the bootstrap for standard errors, however, it is not clear how these will scale with large-scale and high-dimensional datasets.

1.2 Main Contributions

In this study, we propose a Bayesian Cox model based on the flexible counting representation introduced in Andersen & Gill (1982). The advantages of the counting process representation are described in Therneau & Grambsch (2000) and enable to fit models with time-varying covariates, complex censoring and truncation patterns, multiple time-scales, multiple events, and marginal or conditional models for correlated data. By combining our model with a reweighting of the log-likelihood and stochastic variational inference (SVI), we can decouple the sample size from our estimation procedure, enabling inference via subsampling of the full data and hence analyzing data that cannot fit into memory at once e.g. EHR. By using a sparsity inducing prior, we can also do inference in high-dimensional data. Our approach enables the inclusion of time-varying covariates measured at irregularly spaced time points, as well as viable uncertainty estimates for large-scale and high-dimensional datasets.
The main contributions of this paper are:

- Large-scale implementation of a Bayesian version of Cox’s partial likelihood based on a counting process representation.
- Reweighting of Cox’s partial log-likelihood for subsampled data to enable stochastic variational inference
- Viable uncertainty estimates for large-scale and high-dimensional datasets.
- Simulation algorithm for discrete event times with time-varying covariates resembling EHR.
- Application of the proposed method to myocardial infarction in the UK Biobank. This comprises a Cox regression for \( \approx 400,000 \) individuals with \( \approx 1.5 \) million observations and \( \approx 1,000 \) time-varying covariates.

The rest of the paper is structured as follows:

In Section 2 we describe the method in detail. A comprehensive simulation study resembling EHR data with a particular focus on time-varying covariates is provided in Section 3. A modified algorithm to efficiently simulate large-scale and high-dimensional event time data is described in Section 3.1. The results of the simulation are given in Section 3.3 and 3.4, where we compare our method with standard Cox model implementations from R-Survival and R-glmnet. Then, the method is evaluated on real-world applications in Section 4. Section 4.1 shows the model performance on some standard datasets provided in the R-Survival package. Section 4.2 is a small case study on the risk of myocardial infarction in the UK-Biobank. Section 5 discusses some limitations and future directions while Section 6 concludes the paper.
2 Methods

A typical analysis with EHR comprises the emergence of a disease in the population, and one wants to understand the effect of other comorbidities and potential biomarkers on it. The main interest is the time passed until the onset of a particular event, in this case, a disease, denoted here as the event times $T_i$ for an individual $i$.

Generally, one wants to understand how covariates may increase or decrease the risk of the event, consequently accelerating or decelerating the event times. The covariates are allowed to vary over time e.g. the evolving set of diseases an individual acquires or updated laboratory measurements, denoted as a column vector $\mathbf{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))^T$ of the $p$-covariate processes.

Typically, we have incomplete data, like censoring, truncation, or filtering. Here, we assume for simplicity independent right censoring, e.g. physical time limit of observations in EHR. The proposed method, however, can readily be used in conjunction with more complex missing data patterns. Formally, we have an individual censoring time $U_i$ such that we only observe $\tilde{T}_i = \min\{T_i, U_i\}$.

The observed data for $n$ individuals is then $\mathbf{D} = \{(\tilde{T}_i, \mathbf{X}_i(t)) : i = 1, 2, \ldots, n\}$. Note that we do know if $\tilde{T}_i$ is an event or censoring time.

Some important concepts from time-to-event analysis, assuming continuous time, are the survival function $S(t) = \mathcal{P}(T > t)$, i.e. the probability that the event has not happened by $t$, and the hazard function $\alpha(t) = \lim_{\Delta t \to 0} (\Delta t)^{-1} \mathcal{P}(T \leq t + \Delta t | T > t) = f(t)S(t)^{-1}$, i.e. the instantaneous risk of observing an event given the event has not happened yet, where $f(t)$ is the density function. The cumulative hazard is given as $A(t) = \int_0^t \alpha(u)du$, and is linked to the survival function through $S(t) = \exp(-A(t))$. 
2.1 Cox’s partial likelihood based on counting processes

For the observational model we define a multivariate counting process for the \( n \) individuals in the EHR as \( N(t) = (N_1(t), N_2(t), \ldots, N_n(t))^\top \), with individual counting process \( N_i(t), \quad i = 1, 2, \ldots, n \), counting the occurrences of events for individual \( i \) with jump sizes \( \Delta N = 1 \), assuming that no two processes jump simultaneously at \( t \) (no ties).

In principle \( N_i(t) \) can count multiple events e.g. recurrent cancer diagnoses, however, we treat it here as if only a single event occurs, without loss of generality.

The intensity process, the rate at which we expect events to happen, for \( N_i(t) \) is linked to the hazard function under the Cox model by

\[
\lambda^{\gamma, \theta}(t) = Y_i(t) \alpha^{\gamma, \theta}(t) = Y_i(t) \alpha_0^{\gamma}(t) \exp[\theta^\top X_i(t)]
\]

where \( Y_i(t) \) is a predictable process, describing whether individual \( i \) is at risk of observing the event at \( t \).

The baseline hazard \( \alpha_0^{\gamma}(t) \), defined by the possibly infinite-dimensional parameter \( \gamma \), determines the rate at which events occur without any covariate effects.

The \( p \)-regression parameter vector \( \theta \) describes the effect of the time-varying covariates, as they act multiplicatively on the baseline hazard.

The likelihood \( \mathcal{L} \) for independent right-censored event times is the combined effect of the density function for individuals observing events and the survival function for censored individuals, which can na"ively be written by means of the hazard and survival function as

\[
\mathcal{L} = \prod_{i=1}^n f_i^{\delta_i} S_i^{1-\delta_i} = \prod_{i=1}^n \alpha_i^{\delta_i} S_i,
\]

where \( \delta_i \) is an indicator whether the event has happened.

The likelihood for \( \gamma \) and \( \theta \) under the counting process representation of the Cox model follows as

\[
\mathcal{L}(D|\theta, \gamma) = \prod_{t \in T} \left( \prod_{i=1}^n (\alpha^{\gamma}(t) \exp[\theta^\top X_i(t)])^{\Delta N_i(t)} \right) \exp \left[ - \int_0^t \sum_{k=1}^n Y_k(u) \exp[\theta^\top X_k(u)] dA_0^{\gamma}(u) \right],
\]

(1)
where \( T \) is the set of event times and \( A_0^\gamma(t) \) is the cumulative baseline hazard. As we are mainly interested in \( \theta \), it is possible to derive a profile likelihood in the sense of

\[
L(D|\theta) = \max_{\gamma} L(D|\theta, \gamma) = \prod_{t \in T} \prod_i \left( \frac{\exp \left[ \theta^\top X_i(t) \right]}{\sum_{k=1}^n Y_k(t) \exp [\theta^\top X_k(t)]} \right)^\Delta N_i(t),
\]

which is equivalent to the partial likelihood specified by Cox (1972, 1975). For details on how to derive the profile/partial likelihood see Andersen & Gill (1982). The log-likelihood follows as

\[
\log L(D|\theta) = \sum_i \int_0^t \theta^\top X_i(u) dN_i(u) - \int_0^t \log \left( \sum_{k=1}^n Y_k(u) \exp [\theta^\top X_k(u)] \right) dN(u).
\]

For a more detailed derivation of Cox’s partial likelihood under a counting process representation we refer to Andersen et al. (1993).

### 2.2 Reweighting of Cox’s partial likelihood

As can be seen in Equation (2) the partial likelihood cannot be factorized over the data as the denominator contains a sum over the full data. Hence, to use subsamples \( D^* \ll D \) for inference in large-scale datasets, we need to reweight the likelihood, or more precisely, the log-likelihood from Equation (3), appropriately. We reweight the log-likelihood as it will be the main objective for inference. Precisely, a good reweighting should guarantee that \( E_{\pi^*}(\log L(D^*|\theta)) \approx \log L(D|\theta) \), where \( E_{\pi^*}(D) \) is the expectation over random subsamples of the data. To achieve a good approximation we propose the following adjustments to the partial log-likelihood (3):

\[
\log L(D|\theta) \approx w_1 \sum_i \int_0^t \theta^\top X_i(u) dN_i(u) - w_1 \int_0^t \log \left( \sum_{k=1}^{n^*} Y_k(u) \exp [\theta^\top X_k(u)] \right) dN^*(u),
\]

where \( n^* \ll n \) and \( w_1 = \int_0^t 1 dN(u) (\int_0^t 1 dN^*(u))^{-1} \) i.e. the ratio of events, \( w_2 = n (n^*)^{-1} \) i.e. the ratio of observations, with \( N^* \) representing the multivariate counting process corresponding to the subsampled individuals \( n^* \). Particularly the last approximation i.e. the
denominator of the partial likelihood may introduce a bias. However, we show through simulations in the supplementary materials Section 1 that the possible bias is small for most cases encountered in practical applications.

### 2.3 Bayesian Cox Model

To use the partial likelihood defined in Equation (2) for probabilistic inference further justification is needed as it is not necessarily a likelihood in the classical sense. Kalbfleisch & Prentice (1973) have shown that the partial likelihood can be interpreted as proportional to the marginal distribution of the rankings of the event times, however, the derivation does not include time-varying covariates. Kalbfleisch (1978) provides a Bayesian justification based on an underlying diffuse Gamma process for $A_0^\gamma$, showing that the partial likelihood is the limiting marginal posterior distribution. This approach is further extended by Sinha (2003), encompassing a large class of models, including (external) time-varying covariates. We show through simulations in Section 3 that this provides valid estimates even for discretized internal time-varying covariate processes. Hjort (1990) provides additional justifications for the partial likelihood based on Beta and Dirichlet processes.

We can define the posterior distribution following Hjort (1990) and Sinha (2003) and assuming prior independence between $P(\theta)$ and $P(\gamma)$ as:

$$P(\theta|D) \propto \int P(D|\theta, \gamma)P(\theta)P(\gamma)dA_0^\gamma$$

$$\propto P(D|\theta)P(\theta),$$

where $P(D|\theta, \gamma)$ is taken from Equation (1) and $P(D|\theta)$ from Equation (2).

As prior for $P(\theta)$ we use either the Normal distribution for a moderate number of covariates

$$\theta_l|\sigma \sim \text{Normal}(0, \sigma^2), \quad l = 1, 2, \ldots, p$$

or in the case of many covariates and high-dimensional data, the Student-$t_\nu$ distribution
with \( \nu \) degrees of freedom to induce sparsity
\[
\theta_l | \nu, s \sim \text{Student-}t_\nu(0, s^2), \quad l = 1, 2, \ldots, p.
\]

### 2.4 Variational Inference

While it is possible to design an efficient accept/reject algorithm to do MCMC, we recast the inference problem as optimization. Precisely we define a family \( \mathcal{G} \) of variational distributions and try to find the member \( Q(\theta) \) of the family that closely approximates the posterior distribution from Equation (5). Closeness is measured as the Kullback-Leibner (KL-) divergence
\[
Q^*(\theta) = \underset{Q(\theta) \in \mathcal{G}}{\text{argmin}} \ KL(Q(\theta) \ || \ P(\theta|D)) = \underset{Q(\theta) \in \mathcal{G}}{\text{argmin}} \int \log \frac{dQ(\theta)}{dP(\theta)} dQ(\theta).
\]
Generally, it is not possible to optimize the KL-divergence directly, instead we maximize the evidence lower bound (ELBO)
\[
\text{ELBO}(Q) = \mathbb{E}_Q[\log \mathcal{L}(D|\theta)] + \mathbb{E}_Q[\log P(\theta)] - \mathbb{E}_Q[\log Q(\theta)],
\]
where \( \log \mathcal{L}(D|\theta) \) is our reweighted version from Equation (4).

Andersen & Gill (1982) have shown that the estimator of \( \theta \) under the Cox model is asymptotically Normal, hence we propose a Multivariate Normal family for \( \mathcal{G} \) when the number of covariates is moderate, which should give a reasonable approximation, especially for large-scale settings.

For a large number of covariates and high-dimensional data we use a lower rank approximation of the covariance matrix, in the form \( \Sigma = W^T W + I \) where, \( I \) denotes identity matrix, \( W \in \mathbb{R}^{R \times P} \), \( R \) is is the number of ranks and \( P \) is the number of covariates. The low-rank approximation for high-dimensional data is experimental. However, we show through simulations in Section 3.4 that this provides a good approximation. Various methods have been proposed to handle difficult variational objectives by replacing the expectation in the
ELBO through Monte Carlo (MC) samples and optimization via automatic differentiation. For details on some algorithms see Ranganath et al. (2014) and Kucukelbir et al. (2017). A detailed survey of MC-based gradient estimation can be found in Mohamed et al. (2019). We use automatic differentiation variational inference (ADVI) following Kucukelbir et al. (2017) for the optimization of the ELBO, with a random subsample $D^*$ at each optimization step.

3 Simulation

The overall setting for the simulations is the occurrence of a disease in the general population and the association of time-varying covariates with it, emulating a typical scenario encountered when analyzing EHR. Event times are generated based on the Cox proportional hazards model and are measured as age in days, ranging from birth to 80 years. Covariates vary randomly over time at discrete time steps and are drawn from either a Normal or Bernoulli distribution, resembling laboratory measurements or disease status. The baseline hazard reflects typical age-sex adjusted distributions of disease with an initial increase and a decrease in older ages. Individuals are followed from birth until either censoring or an event occurs. Censoring is independent of the data generating process.

3.1 Simulation of event times via the Binomial model

Our proposed algorithm for simulation of event-times follows along the lines of Sylvestre & Abrahamowicz (2008) and is based on the discretized partial likelihood defined by Cox (1972) as

$$\frac{\alpha(t, X)dt}{1 - \alpha(t, X)dt} = \exp \left[ \theta^\top X(t) \right] \frac{\alpha_0(t)dt}{1 - \alpha_0(t)dt}.$$
Based on this, one can define a Binomial model with the probability of an event at time $t$ as

$$p(t) = (1 + \exp \left[ \log(\alpha_0(t)) - \log(1 - \alpha_0(t)) + \theta^\top X(t) \right])^{-1}. \quad (6)$$

For a given simulation, we draw a sample of the underlying baseline hazard. Per individual, we simulate the $p$-covariate processes and then move successively through time $t = 1, 2, \ldots c$ where $c$ is the individual censoring time $c \sim \text{Uniform}(1000, 30000)$. For a given $t$ we evaluate the individual probability of an event as in Equation (6) conditional on the state of the baseline hazard and the covariate processes at $t$ until either an event occurs or the censoring time is reached. The covariate processes up to the final time are then retrieved and kept for the subsequent analysis. A schematic representation of the simulation can be seen in Figure 1.

For the simulations presented in this study, the time between jumps in the covariate processes for each individual is distributed as Gamma$(4, 500)$ rounded to the nearest integer. At every jump, the covariate values are randomly drawn from either a Bernoulli$(0.2)$ distribution or a Normal$(0, 1)$ distribution for binary and quantitative measurements, respectively. We treat the covariates as constant between the jump times. The baseline hazard has successive increments of Gamma$(2, 1)$ until age 40, after which increments increase to Gamma$(1, 10)$ until age 70, after which the baseline hazard decreases in increments of Gamma$(1, 5)$. The time between increments is distributed as Gamma$(4, 200)$ rounded to the nearest integer. The sampled baseline hazard can then be scaled up or down to reflect the appropriate rate of events/censorship. The resulting shape for the baseline hazard is similar to the depiction in the schematic. An example output for two individuals in the resulting long-format can be seen in Table 1.
Figure 1: A schematic overview for discrete failure time simulation. The upper panel shows the rough shape of the baseline hazard $\alpha_0$. The middle panel represents for an individual the corresponding linear effect of the time-varying covariate processes $X_i(t)$ and the parameters $\theta$. The dotted horizontal line shows an increasing hazard $> 0$ or a decreasing hazard $< 0$. The lower panel depicts the combined effect of baseline hazard and covariate effects. The dotted vertical line represents jump times of the baseline hazard and the dashed vertical line the jump times for the covariate processes. The probability of an event $p$ is the combination of both processes with jumps at either of the two and is evaluated for each $dt$ until either censoring or an event happens as indicated by the solid vertical line in the lower panel.
| start | stop  | event | X1  | X2  | X3  | X4  | X5  | X6  |
|-------|-------|-------|-----|-----|-----|-----|-----|-----|
| 0     | 1000  | 0     | 0   | 0   | 0   | 0   | 0   | 0   |
| 1000  | 3235  | 0     | 1   | 0   | 0   | -1.22 | -0.3 | 0.61 |
| 3235  | 6671  | 0     | 0   | 0   | 1   | 0.14 | 0.32 | 0.61 |
| 6671  | 8551  | 0     | 0   | 0   | 0   | -0.76 | -0.2 | 0.65 |
| 8551  | 10146 | 0     | 0   | 1   | 1   | -0.19 | -0.52 | 0.13 |
| 10146 | 11107 | 1     | 0   | 1   | 0   | 1.55 | 0.85 | 1.84 |
| 0     | 3693  | 0     | 0   | 0   | 0   | 0   | 0   | 0   |
| 3693  | 5421  | 0     | 0   | 0   | 0   | -0.42 | -1.87 | 0.34 |
| 5421  | 7291  | 0     | 0   | 0   | 0   | 0.64 | -2.43 | 0.51 |
| 7291  | 9852  | 0     | 0   | 0   | 0   | 0.09 | 0.03 | 0.71 |
| 9852  | 10004 | 0     | 0   | 0   | 0   | -1.23 | 0.76 | -0.03 |

Table 1: Example output for a simulation draw of 2 individuals and 11 observations with 3 binary and 3 quantitative covariates in long format.

### 3.2 Log-likelihood approximation through reweighting

First, we wanted to evaluate the subsampled approximation of the full log-likelihood through the reweighting described in Equation (4), and particularly, how it compares to a more naïve approximation that would simply scale \( \log L \) by the relative number of observations between the full data and the batch size. The data has been generated as:

**Simulation:** 1000 individuals (\( \approx 6111 \) observations), \( \approx 0.65 \) censorship\(^1\), 6 covariates - 3 binary & 3 continuous, \( \theta_l \sim \text{Uniform}(-1.5, 1.5) \), 1024 batch size\(^2\), \( \mathcal{P}(\hat{\theta}_l) \sim \text{Normal}(0, 1) \).

The results can be seen in Figure 2. The reweighting forms an almost unbiased, symmetric distribution around the full log-likelihood throughout the training process and is substantially better than a naïve approximation based on the batch sizes itself. Additionally, we evaluate the performance of our reweighting for different scenarios based on different levels of observations, batch size, censorship, number of covariates and hazard in Section 1 of the supplementary materials. Overall the approximation of the log-likelihood is very good with almost no error on average.

---

\(^1\)Based on individuals throughout the paper  
\(^2\)Based on observations if not stated otherwise
3.3 Standard case

The standard case simulation reflects the typical data structure most often encountered in much of applied medical research, with a small number of time-varying covariates and a moderately sized population, such that $N \gg I \gg P$, i.e. observations $\gg$ individuals $\gg$ covariates. This setting allows us to assess individual parameter estimates based on bias, efficiency, and coverage. Overall we run two separate simulations for different parameter values and varying censorship. The specific settings for the simulations are:

**Simulation 1**: 200 runs, 1000 individuals ($\approx$7630 observations), $\approx$0.78 censorship, 6 covariates - 3 binary & 3 continuous, $\theta = (-0.9, 0.2, 0.0, -0.4, 1.1, 0.0)^\top$.

**Simulation 2**: 200 runs, 2000 individuals ($\approx$16310 observations), $\approx$0.98 censorship, 6 covariates - 3 binary & 3 continuous, $\theta = (0.8, -0.5, 0.0, -0.7, 1.0, 0.0)^\top$. 

Figure 2: Log-likelihood approximation along training. On the left-hand side, we can see the full log-likelihood evaluated at each training step in black, the corresponding reweighted log-likelihood evaluations in dark grey and a naïve approximation in light grey. The right-hand side shows the distributions of the approximations around the full-likelihood evaluated at the last training step for repeated samples.
| θ  | ˆθ  | σ̂θ  | RMSE | HPD_{95%} | Coverage_{95%} | ˆθ  | σ̂θ  | RMSE | CI_{95%} | Coverage_{95%} |
|----|------|-------|------|-----------|---------------|------|-------|------|---------|---------------|
| -0.9 | -0.89 | 0.22 | 0.22 | 0.98 | 0.94 | -0.92 | 0.23 | 0.23 | 0.91 | 0.94 |
| 0.2 | 0.18  | 0.17 | 0.17 | 0.74 | 0.98 | 0.2   | 0.16 | 0.16 | 0.64 | 0.97 |
| 0   | -0.04 | 0.18 | 0.18 | 0.78 | 0.96 | -0.02 | 0.18 | 0.18 | 0.68 | 0.94 |
| -0.4 | -0.39 | 0.07 | 0.07 | 0.32 | 0.96 | -0.39 | 0.06 | 0.07 | 0.27 | 0.96 |
| 1.1 | 1.09  | 0.08 | 0.08 | 0.34 | 0.96 | 1.1   | 0.07 | 0.07 | 0.28 | 0.96 |
| 0   | 0.01  | 0.07 | 0.07 | 0.31 | 0.98 | 0.01  | 0.07 | 0.07 | 0.27 | 0.94 |
| 0.8 | 0.74  | 0.2  | 0.21 | 0.85 | 0.96 | 0.79  | 0.19 | 0.19 | 0.72 | 0.93 |
| -0.5 | -0.52 | 0.25 | 0.25 | 1.11 | 0.96 | -0.49 | 0.24 | 0.24 | 1.01 | 0.98 |
| 0   | -0.05 | 0.23 | 0.23 | 0.99 | 0.96 | -0.01 | 0.23 | 0.23 | 0.87 | 0.94 |
| -0.7 | -0.69 | 0.11 | 0.11 | 0.41 | 0.92 | -0.69 | 0.1  | 0.1  | 0.35 | 0.9 |
| 1   | 0.99  | 0.09 | 0.09 | 0.43 | 0.96 | 0.99  | 0.08 | 0.08 | 0.35 | 0.97 |
| 0   | 0     | 0.09 | 0.09 | 0.4  | 0.94 | 0     | 0.09 | 0.09 | 0.34 | 0.97 |

Notes: RMSE - Root mean squared error, HPD - Highest posterior density, CI - Confidence interval

Table 2: Simulation results for standard case setting. The upper panel shows the results for the first simulation and the lower panel represents the second simulation. The first 3 estimates correspond to the binary covariates, the last 3 to the quantitative covariates.

We use a Multivariate Normal family for variational inference and $\mathcal{P}(\hat{\theta}_l) \sim \text{Normal}(0, 1)$ as prior. The results for the simulation are shown in Table 2. We provide results for the Cox model fitted with the implementation $R$-$\text{Survival}$ in *Therneau & Lumley* (2015) on the full data and our implementation $\text{ProbCox}$ on subsamples. The results for different batch sizes can be found in the supplementary materials Section 2. The average parameter estimates $\hat{\theta}$ are consistent across the different settings and stable for different batch sizes. The variability of the estimates across the simulations measured as the average standard deviation $\sigma_{\hat{\theta}}$, as well as the root mean squared error RMSE, are comparable to $R$-$\text{Survival}$. The coverage is close to the nominal rate with the average highest posterior density interval length $\text{HPD}_{95%}$ similar in size to the confidence interval length from $R$-$\text{Survival}$, increasing
with smaller batches, reflecting the additional approximation error through subsampling. Overall we see comparable performance between our implementation trained on subsamples, and \textit{R-Survival} fitted on the full data.

### 3.4 High-dimensional case

With the number of variables measured on individuals continuously increasing the need for models that can handle large numbers of covariates and high-dimensional data $P \gg N \gg I$ is paramount. Particularly, when analyzing EHR the subsampled data for each evaluation can effectively be very high-dimensional, even though the EHR itself is generally of the structure $N \gg I \gg P$. Hence, a method that can scale to EHR will also need to be able to handle high-dimensional data. Therefore, we evaluated our method on a moderately sized high-dimensional dataset. The simulation has been conducted as follows:

**Simulation**: 200 runs, 1000 individuals ($\approx 7410$ observations), $\approx 0.72$ censorship, 10000 covariates - 5000 binary & 5000 continuous, $\theta_{1:20} \sim \text{Normal}(0, 0.75^2)$, $\theta_{21:10000} = 0$. Parameters are equally split between binary and continuous covariates.

The results for the simulation can be seen in Table 3. We provide a comparison of our proposed method to the Cox model with Lasso penalty as implemented in \textit{R-glmnet} by Friedman et al. (2010) and Simon et al. (2011). The $\lambda$ parameter (strength of penalization) for \textit{R-glmnet} has been chosen by 3-fold cross-validation on the concordance index for $\lambda = 0.01, 0.0115, \ldots, 0.02$. We have evaluated the full Lasso path for $\lambda$ on a single simulation run and the sequence captured most of the relevant range. For \textit{R-glmnet} we show results for $\lambda_{1se}$ as it provides the best parsimonious identification.

We use the low rank approximation to the Multivariate Normal family for variational inference and $\mathcal{P}(\hat{\theta}_l) \sim \text{Student-t}(\nu = 1, s = 0.001)$ as prior. Summaries are computed over the set of estimates that have been identified in the respective simulation runs. For our method a parameter is na"ively classified as identified if the $HPD_{95\%}$ does not contain zero.
| $\theta$ | $\hat{\theta}$ | $\sigma_\theta$ | RMSE | HPD$_{95\%}$ | Coverage$_{95\%}$ | $p_{|\theta|>0}$ |
|---------|--------------|--------------|-------|--------------|-----------------|----------------|
| -0.71   | -0.78        | 0.16         | 0.17  | 0.76         | 0.97            | 0.29           |
| 1.31    | 1.25         | 0.14         | 0.15  | 0.51         | 0.92            | 1              |
| 1.37    | 1.33         | 0.14         | 0.15  | 0.5          | 0.92            | 1              |
| 0.91    | 0.85         | 0.15         | 0.16  | 0.54         | 0.92            | 0.92           |
| 0.4     | 0.53         | 0.06         | 0.14  | 0.56         | 1               | 0.06           |
| -0.19   | 0            | 0.04         | 0.19  | 0.02         | 0               | 0              |
| 0.99    | 0.95         | 0.14         | 0.14  | 0.53         | 0.94            | 0.98           |
| 1.1     | 1.06         | 0.15         | 0.16  | 0.52         | 0.92            | 1              |
| -1.36   | -1.3         | 0.24         | 0.24  | 0.83         | 0.92            | 0.95           |
| -1.16   | -1.1         | 0.2          | 0.21  | 0.79         | 0.93            | 0.87           |
| -0.83   | -0.82        | 0.08         | 0.08  | 0.25         | 0.94            | 1              |
| 0.1     | 0.19         | 0.04         | 0.1   | 0.26         | 0.85            | 0.06           |
| 0.26    | 0.26         | 0.07         | 0.07  | 0.27         | 0.97            | 0.7            |
| 0.82    | 0.82         | 0.09         | 0.09  | 0.25         | 0.92            | 1              |
| -0.53   | -0.51        | 0.08         | 0.08  | 0.25         | 0.94            | 1              |
| -0.4    | -0.38        | 0.08         | 0.08  | 0.25         | 0.9             | 1              |
| 0.5     | 0.49         | 0.08         | 0.08  | 0.25         | 0.94            | 1              |
| -0.35   | -0.33        | 0.07         | 0.08  | 0.26         | 0.94            | 0.96           |
| -0.09   | -0.22        | 0.07         | 0.14  | 0.28         | 0.67            | 0.04           |
| -0.05   | 0            | 0.01         | 0.05  | 0.02         | 0.05            | 0              |

Notes: RMSE - Root mean squared error, HPD - Highest posterior density, $p_{|\theta|>0}$ - probability of identification

Table 3: Simulation results for high-dimensional case - non-zero parameters only. The first ten rows correspond to the non-zero binary covariates and the last ten rows to the non-zero continuous covariates. *R-glmnet* estimates for the standard error are not available.

For *R-glmnet* a parameter is identified, if the estimate is not equal to zero. The rate of identification for the parameters over all simulations is given in the last column of Table 3 for each method, respectively.

The results for *R-glmnet* with $\lambda_{\text{min}}$ and our method with different batch sizes and ranks can be found in the supplementary materials Section 3.
Figure 3: Simulation results for the high-dimensional case. The light grey dots are parameter estimates from individual simulation runs while the black stars represent the average estimate over all simulations. The black bars are the average $HPD_{95\%}$ bounds. The dotted line corresponds to the diagonal and indicates correct identification of the true parameters.

Generally, the estimates for our method work well for both types of covariates, being on average close to the true parameter values and exhibiting good coverage overall. The results are robust across the rank and batch size specification. In comparison to $R$-glmnet, our estimates are much closer to the true parameters, as expected since the Lasso deploys a strong penalization overall. For identification both methods perform equally well with our approach identifying on average 16 covariates with a standard error of 3 and 2 falsely identified compared to $R$-glmnet with on average 15 covariates identified, a standard error of 2 and 1 false identification.

A graphical comparison of our method and $R$-glmnet can be seen in Figure 3. Most of our estimates lie on the diagonal enclosed by the average $HPD_{95\%}$ bounds, indicating again a good identification of the true parameters on average with good quantification of
the uncertainty. The estimates for \textit{R-glmnet} are further off from the diagonal. However, \textit{R-glmnet} does generally identify the correct covariates.

### 3.5 Resource comparison

In this section, we provide a small compute-time and memory utilization comparison between our method \textit{ProbCox}, \textit{R-Survival}, and \textit{R-glmnet}. We run two different versions of our method, one that feeds the subsamples from memory and one version that loads each subsample from the hard drive, reflecting the more typical use-case of our method. Both version are run with the low rank approximation (rank=25) to the Multivariate Normal family for variational inference, \( P(\hat{\theta}_l) \sim \text{Student-}t(\nu = 1, s = 0.001) \) as prior, and a batch size of 256 observations. The main characteristics for the simulation can be taken from Figure 4. Covariates are equally split between binary and continuous covariates, of which ten are non-zero with \( \theta_{1:10} \sim \text{Normal}(0, 0.5^2) \). We only run a single run per simulation per method. Therefore, the estimates only serve as a rough guide. The timings and memory utilization for a 4-core machine can be seen in Figure 4.

Overall we can see a good resource utilization of our method. There is an initial overhead, which is offset once the dataset becomes large-scale and high-dimensional. We can also see a clear trade-off between memory usage and compute time. However, the additional compute time levied through loading the subsamples from the hard-drive is neglectable and allows constant memory usage.
Figure 4: Runtime/Memory comparison between ProbCox, ProbCox (Hard drive) \textit{R-Survival}, and \textit{R-glmnet}. ProbCox (Hard drive) loads the relevant batch sample on demand from the hard drive instead of loading the full data into memory.

4 Application

4.1 Small-scale examples

To further evaluate our proposal and compare it with the standard implementations \textit{R-Survival} and \textit{R-glmnet}, we run a couple of small applications on real-world data provided in \textit{R-Survival}. The focus is on comparing the methods rather than on the data analysis.

In total, we compare 5 different datasets. \textit{Colon} are data on trials of adjuvant chemotherapy for colon cancer. \textit{Lung} are data extracted from the North Central Cancer Treatment Group on mortality for advanced lung cancer. \textit{NAFLD} is a large population-based study investigating non-alcoholic fatty liver disease. \textit{Heart} investigates mortality in patients from
the Stanford heart transplant program. *PBCseq* are follow-up laboratory data from the Mayo clinical trial on primary biliary cholangitis and D-penicillamine treatment conducted between 1974 and 1984. For more details on the data and the covariates we refer to the *R-Survival* documentation.

The results for *Colon*, *Lung*, *NAFLD* and *Heart* can be found in the supplementary material Section 4. Our proposed method compares very well, and the parameter and standard error estimates are similar.

The *PBCseq* data contains 312 individuals with 1,945 observations and 725 events, where we treat repeated events as independent for simplicity. For missing covariates, we carry the information from the previous visit forward. Additionally, we construct an artificial high-dimensional dataset based on the *PBCseq* data by appending 2,000 Normal(0,1) distributed covariates. The continuous covariates from the *PBCseq* have been Z-transformed accordingly. We compare our method to *R-Survival* for the original dataset and *R-glmnet* for the high-dimensional dataset. We use a Multivariate Normal family for variational inference with \( \mathcal{P}(\hat{\theta}_l) \sim \text{Normal}(0,1) \) as prior for the original data and the low rank \((R = 20)\) approximation to the Multivariate Normal family with \( \mathcal{P}(\hat{\theta}_l) \sim \text{Student-}t(\nu = 1, s = 0.001) \) in the high-dimensional case. The batch size for both cases is 256 observations. The results can be see in Table 4.

For the original data, our method is again comparable to the implementation in *R-Survival*, with the same covariates identified as having non-zero effects and similar estimates for parameters and uncertainty. For the high dimensional case, we identify similar covariates as *R-glmnet* with them being the non-zero covariates from the original analysis. However, our estimates are closer to the parameter estimates from the original evaluations compared to *R-glmnet*. Overall our method identified 7 non-zero covariates (based on the \( HPD_{95\%} \)) and *R-glmnet* 8 non-zero covariates.
| Var          | ProbCox   | R-Survival | ProbCox - HD | R-glmnet - $\lambda_{min}$ |
|--------------|-----------|------------|--------------|---------------------------|
|              | $\theta$  | $HPD_{90\%}$ | $\theta$  | $CI_{90\%}$ | $\theta$  | $HPD_{90\%}$ | $\theta$  | $CI_{90\%}$ |
| trt          | -0.07     | (-0.22, 0.08) | -0.07 | (-0.22, 0.08) | 0. | (-) | 0. | (-) |
| sex          | 0.58*     | (0.39, 0.78) | 0.59* | (0.39, 0.78) | 0.54* | (0.34, 0.74) | 0.28 | - |
| edema=0      | -0.55*    | (-0.82, -0.28) | -0.55* | (-0.82, -0.28) | -0.3* | (-0.48, -0.11) | -0.22 | - |
| edema=0.5    | -0.3*     | (-0.47, -0.13) | -0.3* | (-0.55, -0.05) | -0. | (-) | 0. | - |
| stage=2      | -0.12     | (-0.59, 0.35) | -0.13 | (-0.66, 0.4) | 0. | (-) | 0. | - |
| stage=3      | -0.03     | (-0.29, 0.23) | -0.07 | (-0.57, 0.43) | 0. | (-) | 0. | - |
| stage=4      | 0.08      | (-0.09, 0.25) | 0.05 | (-0.45, 0.54) | -0. | (-) | 0. | - |
| ascites      | -0.09     | (-0.31, 0.13) | -0.12 | (-0.36, 0.12) | 0. | (-) | 0. | - |
| hepato       | 0.2*      | (0.03, 0.37) | 0.18* | (0.0, 0.36) | 0.17 | (-0.03, 0.37) | 0.05 | - |
| spiders      | 0.08      | (-0.07, 0.24) | 0.07 | (-0.09, 0.24) | 0. | (-) | 0. | - |
| bili         | 0.3*      | (0.24, 0.37) | 0.31* | (0.24, 0.38) | 0.33* | (0.27, 0.4) | 0.3 | - |
| chol         | -0.05     | (-0.12, 0.02) | -0.05 | (-0.12, 0.02) | 0. | (-) | 0. | - |
| albumin      | -0.25*    | (-0.33, -0.17) | -0.26* | (-0.35, -0.17) | -0.29* | (-0.37, -0.21) | -0.2 | - |
| alk.phos     | 0.18*     | (0.11, 0.24) | 0.18* | (0.11, 0.25) | 0.15* | (0.08, 0.22) | 0.04 | - |
| platelet     | 0.03      | (-0.08, 0.14) | -0.01 | (-0.1, 0.08) | 0. | (-) | 0. | - |
| protime      | 0.15*     | (0.09, 0.21) | 0.15* | (0.09, 0.22) | 0.15* | (0.09, 0.22) | 0.12 | - |
| age          | 0.28*     | (0.21, 0.35) | 0.28* | (0.2, 0.36) | 0.28* | (0.2, 0.35) | 0.18 | - |

Harrel’s C: 0.729 0.729 0.741 0.726
Notes: * non-zero effect at 95%, (-) interval indistinguishable from 0

Table 4: Results for the PBCseq application. The first two columns correspond to the original and the last two columns to the high-dimensional analysis.

### 4.2 Myocardial infarction in the UK Biobank

In this section, we showcase a typical example of analysis where we see the greatest utility of our proposed method, as this analysis would not be possible otherwise. With the number of biobanks and population-based health registers continuously increasing, we suspect these types of analyses to become more widespread and relevant. Being able to conduct multivariate association studies at such a large scale provides new possibilities to identify novel relationships, but also to investigate rare disease outcomes. However, the focus of the
presented small case study is on the method and the type of analysis that can be enabled rather than the actual medical association, therefore the presented result should not be over-interpreted from a medical viewpoint.

Introduction

Cardiovascular disease (CVD) is one of the major causes of death worldwide with an immense impact on public health. CVD is the general term for conditions affecting the heart and blood vessels including ischaemic heart disease and stroke. Understanding the risk factor driving the progression of CVD could help practitioners to intervene early or allow individuals to adjust lifestyle choices. One of the earliest and most influential cohorts in this respect, the Framingham heart study, helped to identify some of the most prevalent factors related to CVD, like smoking, cholesterol, blood pressure, etc., and established a widely applied risk score metric with the same name (Dawber et al. 1951, Mahmood et al. 2014). In this case study, we make use of a prospective cohort, the UK Biobank, to analyze common risk factors associated with myocardial infarction (MI), a subset of CVD, and further investigate hospital admission records to identify possible comorbidities that could be indicative of a future MI event.

Data

The UK Biobank (UKB) (Sudlow et al. 2015) is a public large-scale biomedical database established in 2006, with a vast array of information on 502,628 participants, recruited between 2006 and 2010. All participants were between 40-69 years of age at their recruitment date. Some of the information collected on individuals include: questionnaires and face-to-face interviews (covering general medical factors, lifestyle, environmental factors, socioeconomics, etc.), physical measurements, blood and urine assays, prescriptions, and genotypes, as well as linkage to their hospital admission records.
**Cohort:** In our study, we include all participants from the UKB but filter for participants that have missing information on the baseline risk factors and participants that had a CVD-related event before or in the first year of recruitment. Participants that have withdrawn their consent by the date of this study have also been removed from further analysis. The data has been split for model evaluation into a training, validation, and test set, by proportions of 0.7, 0.1, 0.2, respectively. In total, we have 393,464 participants with 1,594,586 interval observations and 6,880 MI events. Participants are left-truncated by the date of their entry and possibly right-censored, the latest at 2020-03-01.

**Covariates:** We extract some baseline risk factors that are generally be considered to influence an individual’s risk of CVD.

1. *Sex* as indicator [male=1].
2. *Smoking* as indicator for current smoking.
3. *Alcohol* as indicator for alcohol consumption of $\geq$ 3 days per week.
4. *Vigorous activity* as an indicator for at least 2 times vigorous activity for at least 10 minutes per week.
5. *LDL* as a quantitative measure for low-density lipoproteins ("bad cholesterol") in mmol/L centered at the median value 3.516.
6. *HDL* as a quantitative measure for high-density lipoproteins ("good cholesterol") in mmol/L centered at the median value 1.4.
7. *Triglyceride* as a quantitative measure in mmol/L centered at the median value 1.482.
8. Body mass index (BMI kg/m$^2$) with 3 indicators:
   - *Underweight* [BMI < 18.5],
   - *Overweight* [25 $\leq$ BMI < 30],
   - *Obese* [BMI $\geq$ 30].
9. Blood pressure with 3 indicators (systolic blood pressure as (sbp) and diastolic blood pressure as (dbp) in mm Hg):
   - *BP elevated* [120 $\leq$ sbp < 130 and dbp < 80],
   - *Hypertension 1* [130 $\leq$ sbp < 140 or 80 $\leq$ dbp < 90],
   - *Hypertension 2* [sbp $\geq$ 140 or dbp $\geq$ 90].
The hospital records are used to extract information on possible comorbidities, precisely any other disease described in the international classification of disease 10th revision (ICD10) up to the 3rd level of specificity. In fact, the hospital records also contain retrospective data and include additional information from general practitioner visits, an individual’s own recollection, and possibly death certificates. The code for obesity (E66) and essential primary hypertension (I10) have been integrated into the corresponding baseline risk factor to avoid double specification. Some of the CVD-related diseases have been removed as they are used as event codes or for censoring. In total, we have 1108 time-varying binary indicators for each ICD10 code representing if an individual ever had the corresponding disease at the corresponding point in time.

Overall, there are 1120 time-varying covariates and sex in our model, with covariates values taken from either the most recent recruitment visit, follow up or entry in the hospital admission records and are assumed to be constant in-between measurements. We transform the timeline from real-time to age in days, as age is a dominant and powerful predictor, in these cases, it can be advisable to use age as the preferable time axis. See Andersen et al. (2021) for general guidelines.

**Outcomes:** The outcome of the study is a myocardial infarction defined as an ICD10 code in I21-I24. Other CVD events, ischemic heart disease (I20-I25), cerebrovascular disease (I60-I69), cardiac arrest (I46), heart failure (I50), and transient cerebral ischaemic attack (G45) have been used additionally as either exclusion criterion before study entry or as a possible competing event, which we treat here as a censoring instance for simplicity. Of interest is the MI event one year ahead - the last year of observation prior to the event is removed - to avoid identifying factors that might be part of the diagnostic process itself.
Model

For analysis, we use a similar setup as in the high-dimensional simulation in Section 3.4 as we have $\approx 1000$ covariates and would only assume a small fraction to be actually relevant. We estimate a linear effect for the covariates $\theta^\top X(t)$. The prior distribution is $P(\theta_l) \sim \text{Student-t}(\nu = 1, s = 0.001)$, where we use a low rank Multivariate Normal with Rank = 20 as our approximation family for variational inference. The batch size for inference is 8,192 individuals.

Results

Overall the results are in line with many of the reported associations in the literature. The most relevant factors according to our estimation can be seen in Figure 5.

Yusuf et al. (2020) reports a hazard ratio of 1.74 (1.61 - 1.88) for diabetes without sub-categorization on CVD, which agrees with our estimates for unspecified diabetes 1.66 (1.5 - 1.84) and insulin-dependent diabetes 1.73 (1.31 - 2.29). Millett et al. (2018) estimates sex-specific factors for smoking on MI as 3.46 (3.02 - 3.98) and 2.23 (2.03 - 2.44) for female and male, respectively, slightly higher than our estimate of 2.06 (1.91 - 2.23). Mortensen & Nordestgaard (2020) estimates a hazard ratio of 1.34 (1.27 - 1.47) for LDL on MI, again close to our estimates of 1.41 (1.37 - 1.46). Chronic kidney disease has been identified by Hippisley-Cox et al. (2017) as a risk factor for CVD with a hazard ratio of 2.09 (1.87 - 2.34) and 1.94 (1.72 - 2.19), for male and female, respectively. These estimates overlap with our cystic kidney disease and hypertensive renal disease estimates of 1.98 (1.15-3.4) and 1.89 (1.15 - 3.1), respectively. The same study also estimates a hazard ratio for rheumatoid arthritis of 1.24 (1.19 - 1.28), similar to our estimate of 1.29 (1.04 - 1.61). The association between rosacea and MI is less well established, however, Egeberg et al. (2016) estimates a relative risk ratio of 0.75% (0.57 - 1.00) in line with our estimate of 0.6 (0.37 - 0.98). The concordance index for the model on the train, valid, and test split
are 0.72, 0.72, and 0.69, respectively. This type of analysis could be further extended to include more of the available information in the UKB and refine associations, but more importantly, with the number of participants in these cohorts, we can now reveal and study more nuanced outcomes and systematically evaluate multivariate effects.

Figure 5: Forest plot for the hazard exp[θ]. The black dot represents the median parameter estimates with the corresponding HPD_{95%} around it. The upper panel shows the estimates for the baseline risk factors. We include all baseline risk estimates even if they have not been identified. The lower panel shows the parameter estimates for diseases with sufficient evidence of a non-zero effect based on the HPD_{95%} interval.
5 Limitations and future directions

The presented analysis mainly focused on the scalability of the Cox model in high-dimensional and large-scale datasets with time-varying covariates. Further research will be needed to extend the proposed method to some of the challenges that may arise when analyzing EHR. We provide a brief outline of some possible future directions of this research.

A limitation of the proposed analysis is the assumption of proportional hazards and the lack of means to assess. Developing efficient algorithms to scale residuals analysis to these vast amounts of data to evaluate model performance and identify possible deficiencies will be crucial.

Further extending the flexibility of the method by non-linear and time-dependent covariate effects is another factor to consider. However, challenges may arise through the different dimensions on which the effects can act when considering time-varying covariates. The effect may vary by timepoint of measurement (early vs. late pregnancy) and by the progression in time (decaying or increasing effects with time).

While EHR provide a rich data source, the assumption that the measurements are constant in-between time points or are measured without error may introduce a bias. Modeling the longitudinal data jointly with event times would be an alternative approach, however, there are additional inferential and computational challenges.

Competing events naturally arise when analyzing EHR and pose challenges for the interpretation of results. In this study we focused on the cause-specific hazard, however, adopting different approaches like the Fine-Gray model, i.e. subdistribution hazard, may be preferable in some cases while challenging for internal time-varying covariates.

Using SVI rather than MCMC provides substantial compute time advantages at the cost of lacking guarantees for the approximation. The adoption of SVI for more complex models e.g. hierarchical models or frailty models imposes additional challenges as defining a good approximating family will require further research and evaluation. The performance
of our approach in the context of high-dimensional data is promising. Nevertheless, using
the lower rank approximation to the Normal distribution as the variational family will
require further theoretical justifications. Extending SVI to very high-dimensional datasets
e.g genetic association studies or using flexible variational families e.g. normalizing flows
are promising research directions.

6 Concluding remarks

We have proposed a Bayesian version of Cox partial likelihood that can be used on high-
dimensional and large-scale datasets with time-varying covariates. Our method provides
(almost) unbiased estimates with good uncertainty quantification while subsampling data.
This approach enables researchers to extend the Cox model, one of the most widely ap-
plied methods in biomedical research, to new data sources like biobanks and EHR, the
combination of which can provide new insights into our understanding of diseases. Our
implementation enables the joint analysis of tens of thousands of time-varying covariates
for millions of individuals, providing a framework to utilize population-scale EHR. Based
on a counting process representation, our method can easily be extended to a wider variety
of scenarios like recurrent events or more elaborate missing data patterns. As shown in
the case of myocardial infarction, we can replicate many of the results previously reported
in the literature and characterize their joint effects. Extending the analysis by the addi-
tional data typically available in EHR or focusing on less well-studied disease could be an
interesting avenue to follow.
Acknowledgments

AWJ and MG are supported by grant NNF17OC0027594 from the Novo Nordisk Foundation. The data for the UK Biobank was accessed by application 45761.

SUPPLEMENTARY MATERIAL

Python-package probcox: Contains the code to fit the probabilistic cox regression method described in the article. The package also contains all datasets and scripts (except the UKB data) used as examples in the article. (GNU zipped tar file)

Appendix Contains additional tables and results described in the article. (pdf file)

References

Alvares, D., Lázaro, E., Gómez-Rubio, V. & Armero, C. (2021), ‘Bayesian survival analysis with BUGS’, Statistics in Medicine. Publisher: Wiley Online Library.

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (1993), Statistical Models Based on Counting Processes, Springer Series in Statistics, Springer-Verlag, New York. URL: https://www.springer.com/gp/book/9780387945194

Andersen, P. K. & Gill, R. D. (1982), ‘Cox’s Regression Model for Counting Processes: A Large Sample Study’, The Annals of Statistics 10(4), 1100–1120. URL: http://projecteuclid.org/euclid.aos/1176345976

Andersen, P. K., Perme, M. P., Houwelingen, H. C. v., Cook, R. J., Joly, P., Martinussen, T., Taylor, J. M. G., Abrahamowicz, M. & Therneau, T. M. (2021), ‘Analysis of time-to-event for observational studies: Guidance to the use of intensity models’, Statistics in Medicine 40(1), 185–211. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8757. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8757

Clift, A. K., Coupland, C. A., Keogh, R. H., Diaz-Ordaz, K., Williamson, E., Harrison, E. M., Hayward, A., Hemingway, H., Horby, P., Mehta, N. & others (2020), ‘Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study’, bmj 371. Publisher: British Medical Journal Publishing Group.
Cox, D. R. (1972), ‘Regression Models and Life-Tables’, *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2), 187–220. Publisher: [Royal Statistical Society, Wiley].

**URL:** https://www.jstor.org/stable/2985181

Cox, D. R. (1975), ‘Partial Likelihood’, *Biometrika* 62(2), 269–276. Publisher: [Oxford University Press, Biometrika Trust].

**URL:** https://www.jstor.org/stable/2335362

Dawber, T. R., Meadors, G. F. & Moore Jr, F. E. (1951), ‘Epidemiological approaches to heart disease: the Framingham Study’, *American Journal of Public Health and the Nations Health* 41(3), 279–286. Publisher: American Public Health Association.

Egeberg, A., Hansen, P. R., Gislason, G. H. & Thyssen, J. P. (2016), ‘Assessment of the risk of cardiovascular disease in patients with rosacea’, *Journal of the American Academy of Dermatology* 75(2), 336–339. Publisher: Elsevier.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), ‘Regularization Paths for Generalized Linear Models via Coordinate Descent’, *Journal of Statistical Software* 33(1), 1–22.

**URL:** https://www.jstatsoft.org/v33/i01/

Hippisley-Cox, J. & Coupland, C. (2021), ‘Predicting the risk of prostate cancer in asymptomatic men: a cohort study to develop and validate a novel algorithm’, *British Journal of General Practice* 71(706), e364–e371. Publisher: British Journal of General Practice.

Hippisley-Cox, J., Coupland, C. & Brindle, P. (2017), ‘Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study’, *bmj* 357. Publisher: British Medical Journal Publishing Group.

Hjort, N. L. (1990), ‘Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data’, *The Annals of Statistics* 18(3), 1259–1294. Publisher: Institute of Mathematical Statistics.

**URL:** https://www.jstor.org/stable/2242052

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001), *Bayesian Survival Analysis*, Springer Series in Statistics, Springer-Verlag, New York.

**URL:** https://www.springer.com/gp/book/9780387952772

Kalbfleisch, J. D. (1978), ‘Non-Parametric Bayesian Analysis of Survival Time Data’, *Journal of the Royal Statistical Society. Series B (Methodological)* 40(2), 214–221. Publisher: [Royal Statistical Society, Wiley].

**URL:** https://www.jstor.org/stable/2984758

Kalbfleisch, J. D. & Prentice, R. L. (1973), ‘Marginal Likelihoods Based on Cox’s Regression and Life Model’, *Biometrika* 60(2), 267–278. Publisher: [Oxford University Press, Biometrika Trust].

**URL:** https://www.jstor.org/stable/2334538

32
Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D. M. (2017), ‘Automatic Differentiation Variational Inference’, Journal of Machine Learning Research 18(14), 1–45. 
URL: http://jmlr.org/papers/v18/16-107.html

Kvamme, H., Borgan, O. & Scheel, I. (2019), ‘Time-to-Event Prediction with Neural Networks and Cox Regression’, Journal of Machine Learning Research 20(129), 1–30. 
URL: http://jmlr.org/papers/v20/18-424.html

Laud, P. W., Damien, P. & Smith, A. F. M. (1998), Bayesian Nonparametric and Covariate Analysis of Failure Time Data, in D. Dey, P. Müller & D. Sinha, eds, ‘Practical Nonparametric and Semiparametric Bayesian Statistics’, Lecture Notes in Statistics, Springer, New York, NY, pp. 213–225.

Li, R., Chang, C., Justesen, J. M., Tanigawa, Y., Qiang, J., Hastie, T., Rivas, M. A. & Tibshirani, R. (2020), ‘Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank’, Biostatistics (kxaa038).
URL: https://doi.org/10.1093/biostatistics/kxaa038

Mahmood, S. S., Levy, D., Vasan, R. S. & Wang, T. J. (2014), ‘The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective’, The Lancet 383(9921), 999–1008. Publisher: Elsevier.

Millett, E. R. C., Peters, S. A. E. & Woodward, M. (2018), ‘Sex differences in risk factors for myocardial infarction: cohort study of UK Biobank participants’, BMJ p. k4247.
URL: https://www.bmj.com/lookup/doi/10.1136/bmj.k4247

Mittal, S., Madigan, D., Burd, R. S. & Suchard, M. A. (2014), ‘High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis’, Biostatistics (Oxford, England) 15(2), 207–221.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3944969/

Mohamed, S., Rosca, M., Figurnov, M. & Mnih, A. (2019), ‘Monte carlo gradient estimation in machine learning’, arXiv preprint arXiv:1906.10652.

Mortensen, M. B. & Nordestgaard, B. G. (2020), ‘Elevated LDL cholesterol and increased risk of myocardial infarction and atherosclerotic cardiovascular disease in individuals aged 70–100 years: a contemporary primary prevention cohort’, The Lancet 396(10263), 1644–1652.
URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673620322339

Nikooienejad, A., Wang, W. & Johnson, V. E. (2020), ‘Bayesian variable selection for survival data using inverse moment priors’, The annals of applied statistics 14(2), 809. Publisher: NIH Public Access.

Qiou, Z., Ravishanker, N. & Dey, D. K. (1999), ‘Multivariate survival analysis with positive stable frailties’, Biometrics 55(2), 637–644.
Ranganath, R., Gerrish, S. & Blei, D. (2014), Black box variational inference, in ‘Artificial intelligence and statistics’, PMLR, pp. 814–822.

Sharef, E., Strawderman, R. L., Ruppert, D., Cowen, M. & Halasyamani, L. (2010), ‘Bayesian adaptive B-spline estimation in proportional hazards frailty models’, *Electronic Journal of Statistics* **4**(none), 606–642. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

URL: https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-4/issue-none/Bayesian-adaptive-B-spline-estimation-in-proportional-hazards-frailty-models/10.1214/10-EJS566.full

Shin, M., Bhattacharya, A. & Johnson, V. E. (2018), ‘Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings’, *Statistica Sinica* **28**(2), 1053. Publisher: NIH Public Access.

Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2011), ‘Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent’, *Journal of Statistical Software* **39**(5), 1–13.

Sinha, D. (1993), ‘Semiparametric Bayesian Analysis of Multiple Event Time Data’, *Journal of the American Statistical Association* **88**(423), 979–983. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

URL: https://www.jstor.org/stable/2290789

Sinha, D. (2003), ‘A Bayesian justification of Cox’s partial likelihood’, *Biometrika* **90**(3), 629–641.

URL: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/90.3.629

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. & others (2015), ‘UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age’, *Plos med* **12**(3), e1001779. Publisher: Public Library of Science.

Sylvestre, M.-P. & Abrahamowicz, M. (2008), ‘Comparison of algorithms to generate event times conditional on time-dependent covariates’, *Statistics in Medicine* **27**(14), 2618–2634. _eprint:_ https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.3092.

URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3092

Tarkhan, A. & Simon, N. (2020), ‘BigSurvSGD: Big Survival Data Analysis via Stochastic Gradient Descent’, *arXiv:2003.00116 [math, stat]_. arXiv: 2003.00116.

URL: http://arxiv.org/abs/2003.00116

Therneau, T. M. & Grambsch, P. M. (2000), The cox model, in ‘Modeling survival data: extending the Cox model’, Springer, pp. 39–77.

Therneau, T. M. & Lumley, T. (2015), ‘Package ‘survival’, *R Top Doc* **128**(10), 28–33.
Tibshirani, R. (1996), ‘Regression Shrinkage and Selection via the Lasso’, *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1), 267–288. Publisher: [Royal Statistical Society, Wiley].

**URL:** [https://www.jstor.org/stable/2346178](https://www.jstor.org/stable/2346178)

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

Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., Curtis, H. J., Mehrkar, A., Evans, D., Inglesby, P., Cockburn, J., McDonald, H. I., MacKenna, B., Tomlinson, L., Douglas, I. J., Rentsch, C. T., Mathur, R., Wong, A. Y. S., Grieve, R., Harrison, D., Forbes, H., Schultze, A., Croker, R., Parry, J., Hester, F., Harper, S., Perera, R., Evans, S. J. W., Smeeth, L. & Goldacre, B. (2020), ‘Factors associated with COVID-19-related death using OpenSAFELY’, *Nature* 584(7821), 430–436. Number: 7821 Publisher: Nature Publishing Group.

**URL:** [https://www.nature.com/articles/s41586-020-2521-4](https://www.nature.com/articles/s41586-020-2521-4)

Witten, D. M. & Tibshirani, R. (2010), ‘Survival analysis with high-dimensional covariates’, *Statistical methods in medical research* 19(1), 29–51.

**URL:** [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4806549/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4806549/)

Yang, Y. & Zou, H. (2013), ‘A cocktail algorithm for solving the elastic net penalized Cox’s regression in high dimensions’, *Statistics and its Interface* 6(2), 167–173. Publisher: International Press of Boston.

Yusuf, S., Joseph, P., Rangarajan, S., Islam, S., Mente, A., Hystad, P., Brauer, M., Kutty, V. R., Gupta, R., Wielgosz, A., AlHabib, K. F., Dans, A., Lopez-Jaramillo, P., Avezum, A., Lanas, F., Oguz, A., Kruger, I. M., Diaz, R., Yusoff, K., Mony, P., Chifamba, J., Yeates, K., Kelishadi, R., Yusufali, A., Khatib, R., Rahman, O., Zatonska, K., Iqbal, R., Wei, L., Bo, H., Rosengren, A., Kaur, M., Mohan, V., Lear, S. A., Teo, K. K., Leong, D., O’Donnell, M., McKee, M. & Dagenais, G. (2020), ‘Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study’, *The Lancet* 395(10226), 795–808.

**URL:** [https://linkinghub.elsevier.com/retrieve/pii/S0140673619320082](https://linkinghub.elsevier.com/retrieve/pii/S0140673619320082)