Check for updates

OPEN ACCESS

EDITED BY Sebastian Zöllner, University of Michigan, United States

REVIEWED BY

Ghislain Rocheleau, Icahn School of Medicine at Mount Sinai, United States Matthew Zawistowski, University of Michigan, United States Silke Szymczak, University of Lübeck, Germany

*CORRESPONDENCE Anastassia Kolde, ⊠ anastassia.kolde@ut.ee

¹These authors have contributed equally to this work and share first authorship

RECEIVED 26 November 2024 ACCEPTED 15 May 2025 PUBLISHED 25 June 2025

CITATION

Kolde A, Koitmäe M, Käärik M, Möls M and Fischer K (2025) Analysis of follow-up data in large biobank cohorts: a review of methodology. *Front. Genet.* 16:1534726. doi: 10.3389/fgene.2025.1534726

COPYRIGHT

© 2025 Kolde, Koitmäe, Käärik, Möls and Fischer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Analysis of follow-up data in large biobank cohorts: a review of methodology

Anastassia Kolde^{1.2*†}, Merli Koitmäe^{1.2†}, Meelis Käärik¹, Märt Möls¹ and Krista Fischer^{1.2}

¹Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia, ²Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia

This study focuses on key methodological challenges in genome-wide association studies (GWAS) of biobank data with time-to-event outcomes, analyzed using the Cox proportional hazards (CPH) model. We address four primary issues: left-truncation of the data, computational inefficiency of standard model-fitting algorithms, relatedness among individuals, and model misspecification. To manage left-truncation, the common practice is to use age as the timescale, with individuals entering the risk set at their age of recruitment. We assess how this choice of timescale influences bias and statistical power, under realistic GWAS conditions of varying effect sizes and censoring rates. In addition, to alleviate the computational burden typical in large-scale data, we propose and evaluate a two-step martingale residual (MR) approach for high-dimensional CPH modeling. Our results show that the timescale choice has minimal effect on accuracy for small hazard ratios, though using time since birth as the timescale - ignoring recruitment age yields the highest power for association detection. We find that relatedness, when ignored, does not substantially bias effect size estimates, while omitting key covariates introduces significant bias. The two-step MR approach proves to be computationally efficient, retaining power for detecting small effect sizes, making it suitable for large-scale association studies. However, when precise effect size estimates are critical, particularly for moderate or larger effect sizes, we recommend recalculating these estimates using the conventional CPH model, with careful attention to left-truncation and relatedness. These conclusions are drawn from simulations and illustrated with data from the Estonian Biobank cohort.

KEYWORDS

survival analysis, genome-wide association study, populationbased biobank data, martingale residuals, cox proportional hazards model

1 Introduction

As time goes on, the data volume in large-scale population-based biobanks is increasing exponentially. Although the recent decades have seen tremendous increases in sample size, a similarly valuable data expansion results from prolonged follow-up time and the ability to link the -omics databases with incident disease data from electronic health records. Therefore, a large proportion of Genome-Wide Association Studies (GWAS) are mainly focused on discovery of genetic variants associated with the risk of incident diseases. For that purpose, one needs to apply regression modelling methodology that is designated for censored time to event data, rather than using simple methods like linear or logistic regression models (Hughey et al., 2019; Syed et al., 2016; Staley et al., 2017). Here, the Cox Proportional Hazards (CPH) modelling (Cox, 1972) has become a standard in biomedical fields due to its robustness to distributional assumptions and interpretation of parameter estimates in terms of Hazard Ratios (HRs).

Despite its robustness, CPH modelling is not assumption-free. Therefore there is a need for a review of applicability of this method in the context of large population-based biobanks. Our aim is to identify sources of possible biases as well as realistic magnitudes of them in typical GWAS settings.

The most discussed assumption in the context of CPH model is the proportional hazards assumption, stating that the multiplicative effect of a risk factor on the hazard of the outcome event is staying the same throughout the scale of the follow-up time. Recently it has been pointed out that this is rarely true in practice-on the contrary, the hazard ratios are almost always time-varying. Therefore the HR from a CPH model should be interpreted as a weighted average of the true HRs over the follow-up period (Stensrud and Hernán, 2020; Hernán, 2010). This could be easily acceptable in GWAS, unless there is a reason to believe that some genetic variants have a drastically different effect on the risk during different segments of the follow-up time.

Some other, often ignored assumptions are related to the special features of the biobank data. First, we note that the time of recruitment is usually not a relevant baseline timepoint regarding the outcome event (unlike in clinical trials, where follow-up often starts at diagnosis). As the genomic data stays largely constant throughout the lifetime of an individual, date of birth may seem as a logical time origin for a GWAS. Using age at the outcome event as the outcome variable can, however, lead to another problem called left-truncation or immortal time bias (Thiébaut and Bénichou, 2004; Korn et al., 1997), as the analysis is still conducted conditionally on the fact that the individual was alive at the time of recruitment and free of diseases (sometimes including the outcome event) that would have prevented the recruitment. To properly account for left-truncation, one should use methods that use age as timescale, but consider the individual as being at risk only during the time from recruitment until the outcome event (or end of follow-up) (Thiébaut and Bénichou, 2004).

A common feature of biobank cohorts is genetic relatedness of the participants, violating the assumption of independent observations in the sample. For continuous trait GWAS, the use of mixed linear models has been recommended in such cases (Thornton et al., 2012; Yu et al., 2006; Kang et al., 2008; Zhang et al., 2010; Pankratz et al., 2005). Similar approach could be used in survival analysis (mixed effects Cox regression, frailty models) (Dey et al., 2022).

When the outcome event is an incident disease, mortality due to other causes will always be a competing event-censoring the individuals where the follow-up ended due to death, ignores the assumption of independent censoring. In this case, one may consider using a proper model for competing risks (Fine and Gray model). However, when the focus is not on risk prediction, but on parameter estimation, censoring the competing outcomes is still acceptable (Austin et al., 2016; Therneau and Grambsch, 2000).

As the biobank cohorts are mostly not random samples from the population, also other sources of selection bias are likely (Schoeler et al., 2023), that could sometimes be addressed by proper use of sampling weights.

In addition to the biases resulting from sample design, also some computational approaches used in GWAS may become sources of bias. Due to the significant increase in the size of genotyped samples over the past few decades, both in terms of the number of genotyped subjects and the number of genetic variants genotyped or imputed, most of proposed tools for running CPH modeling in GWAS setting have become computationally prohibitive and not easily scalable (Rizvi et al., 2019; Lemieux Perreault et al., 2016; Syed et al., 2017; Gogarten et al., 2012). In some studies, a two-stage approach involving martingale residuals that dramatically reduces the computing time, has been used (Joshi et al., 2016; Timmers et al., 2019). Although it has shown to perform well in some simulations, it is still unclear, whether it creates biases in parameter estimates in some realistic cases.

In classical linear regression with independent covariates, consistent and unbiased estimates for the remaining coefficients can be obtained even if some covariates are omitted. However, this property does not extend to non-linear models, including logistic regression and CPH model neither in randomized or observational setting (Gail et al., 1984; Lagakos and Schoenfeld, 1984; Struthers and Kalbfleisch, 1986; Bretagnolle and Huber-Carol, 1988; Morgan et al., 1986; Lin et al., 2013).

In a CPH model, omitting an important risk factor for the outcome leads to violation of the proportionality assumption with respect to other variables in the model, leading to omitted variable bias, which can significantly distort estimates and conclusions.

In GWAS context one still needs to keep in mind the main task of identifying the potential disease-associated variants in the set of a large number (often more than 20 millions) of genotyped variants. As the focus is on hypothesis testing rather than precise effect estimation, small biases are not a cause of concern, if the nominal Type 1 error is properly controlled after. Therefore, if there is a trade-off between bias and power, a biased estimator may be preferable if it leads to greater power.

The main aim of the present study is to assess the magnitude of bias and power in realistic GWAS settings, where the "naive" CPH model is used, while ignoring left-truncation and/or relatedness of individuals, possibly using the martingale residual approach to speed up the computation. We explore these questions analytically and also by a simulation study, clarifying the need for various bias-reduction measures in GWAS settings.

Finally we also address the option to combine participant genotypes and parental outcome data, when the biobank cohort is relatively recent and the number of events still low (especially for mortality outcomes). Clearly, the estimated HRs in this case will not be unbiased. We derive the expression for the bias analytically and demonstrate the performace of this approach in a small-scale simulation. All simulations and analyses are performed using R software (R Core Team, 2023).

2 Sources of bias

2.1 Timescale choice

2.1.1 The effect of timescale on the likelihood for the cox proportional hazards model

Suppose T is the time to event variable of interest. Here and hereafter we are considering the Cox proportional hazards (CPH) model defined as

$$h(t|\mathbf{X}) = h_0(t)e^{\mathbf{X}\boldsymbol{\beta}},$$

where $h_0(t)$ is the baseline hazard function at time T = t, **X** is the matrix of covariates and β is the vector of the parameters.

To estimate the parameter $\psi = e^{\beta}$ using the CPH model, one needs to find the value of β that maximizes the partial likelihood function:

$$L_{c}(\boldsymbol{\beta}) = \prod_{j=1}^{r} L_{j}(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(x_{(j)}\boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_{(j)})} \exp(x_{l}\boldsymbol{\beta})},$$

where $x_{(j)}$ is the value of *X* for an individual who died at the *j*th observed death time $t_{(j)}$, x_l are the values of *X* for the individuals in the risk set at $t_{(j)}$ and *r* is the number of events. Note that the risk set $\mathcal{R}(t_{(j)})$ consists of individuals observed to survive up to $t_{(j)}$.

The choice of timescale determines the definition of the time variable T, with important implications for how survival time and risk sets are interpreted. In what follows, we use death as the event of interest for concreteness, but the event could be any relevant outcome (e.g., diagnosis, recovery, or relapse).

- 1. When the timescale is time since birth, *T* is defined as age at death, with $t_{(1)}$ being the survival time of the youngest individual at death, etc. The risk set $\mathcal{R}(t_{(j)})$ contains individuals who died at older age than the individual dying at age $t_{(j)}$. The size of the risk set decreases with increasing *t* as each deceased or censored individual is removed from the risk set and no additional individuals are added. This time scale ignores the recruitment and follow-up times-the risk set at $t_{(j)}$ may contain individuals who are recruited to the study at a later age than $t_{(j)}$.
- 2. When *T* is time since recruitment, then the death times $t_{(j)}$ are ordered according to the time spent under follow-up, regardless of the actual age of the participant. The risk set $\mathcal{R}(t_{(j)})$ contains individuals who, since recruitment have been followed for a longer period than t_j . Thus, the size of risk set here is also a decreasing function of *t*.
- 3. With age as timescale, *T* is the age at death as in case 1. However, the risk set only contains the individuals who are under follow-up at the age $t_{(j)}$. Therefore the size of the risk set is not necessarily decreasing, as an individual who is recruited at age *t* and dies or is censored at age t_* is only included in the risk sets $\mathcal{R}(t_{(j)})$ where $t \le t_{(j)} < t_*$.

For example, consider individual A who was born in 1940, recruited into the study in 2010 at age 70, and died in 2015 at age 75. Under time since birth as the timescale, A's event time is 75 years. Under time since recruitment, the event time is 5 years. When age is used as the timescale, A contributes to the risk sets only between ages 70 and 75, the period during which they are under observation.

In the analysis of a birth cohort, all three apporaches are equivalent as in this case time since recruitment is equivalent to time since birth. In studies where recruitment time is a meaningful time origin (such as time at a diagnosis or start of treatment), it is natural to use time since recruitment as timescale. In biobank cohorts it would be more natural to use time since birth as timescale.

2.1.2 Potential bias with time since birth as timescale

Obviously, for unbiased estimation the risk set $\mathcal{R}(t_{(j)})$ should be a random sample from the corresponding risk set in the population. However, when time of birth is used as the time origin (case 1) the risk set at *t* is partly selected conditionally on events (recruitments) that occur after *t* and thus also conditionally on the fact that the individuals survive (and have no outcome events) before their potential recruitment time (say, time $t + \delta$). As a result, low-risk individuals will be over-represented in that risk set (as they have more chance to survive up to time $t + \delta$), leading to overestimation of the hazard ratio.

Still, while using time of birth as time origin and ignoring left truncation, — and assuming that age at recruitment is independent of the covariates X — no bias is anticipated under the null hypothesis ($\psi = 1$, equivalently $\beta = 0$). Therefore the question remains, whether the approach of using age as timescale and time of birth as time origin could still be valid or even preferred for hypothesis testing due to potentially better power.

2.1.3 Potential bias with time since recruitment as timescale

A standard "textbook" approach would be to use time since recruitment (case 2) as timescale, while the effect of age is accounted for by proper adjustment. Korn et al. (1997) have determined two sufficient conditions-when either one is satisfied, the age-adjusted CPH models with time since recruitment as timescale and CPH models with age as timescale give the same results:

- 1. The baseline hazard function $h_0(t)$ can be presented as an exponential function: $h_0(t) = c \cdot \exp(\gamma t)$ for some c > 0 and γ .
- 2. The covariate of interest and age at recruitment are statistically independent.

If neither of the two conditions is fulfilled, Korn et al. suggest using CPH models with age as timescale, as they believe the outcomes to change more as a function of age rather than a function of time since recruitment.

The first condition holds when hazard of the outcome is expected to increase rapidly as a function of age, following a Gompertz distribution (often appropriate for human mortality data).

Korn et al. did not provide a formal proof for the second condition, whereas Thiébaut and Bénichou (2004) found that mismodeling of age as an adjustment factor in a follow-updependent CPH model rather than as a timescale could also result in bias, on the contrary to the reasoning by Korn et al. Other authors have shown that bias can be detected even when a variable independent from the variable of interest has been omitted from the CPH model (Gail et al., 1984; Lagakos and Schoenfeld, 1984; Struthers and Kalbfleisch, 1986; Bretagnolle and Huber-Carol, 1988). One can argue that while modelling all-cause mortality, the second condition basically states that the covariate of interest cannot affect mortality as the distribution of this covariate would otherwise change with age, making them dependent. The distribution stays invariable, when the covariate does not affect mortality.

Thiebaut et al. also suggest using age as timescale rather than time since recruitment, as the underlying mechanisms of these



models are different. They point out that usually the time when subject comes under observation does not coincide with the time when the subject becomes at risk for the outcome of interest. This is especially true in the biobank context.

Again, our question is related to the practical implementation of these findings in the context of GWAS–what is the effect of timescale choice in a range of realistic settings on bias in parameter estimates as well as on power to detect an association. We will try to shed some light on these questions using a simulation study.

2.2 Dependent observations

Independence of observations is a central assumption in most modeling approaches, including the CPH model. However, in population-based biobanks it is common to encounter genetically related individuals, which violates this assumption and can introduce biases. A general approach to address population structure in GWAS is to add principal components (PCs) as covariates (Novembre and Stephens, 2008; Price et al., 2006). While PCs primarily address population stratification, they also might help in identifying and accounting for cryptic relatedness within the sample. However, specialized methods are preferable for explicitly adjusting for related individuals alongside population stratification. These methods include using a kinship matrix (Thornton et al., 2012), linear mixed models (Yu et al., 2006; Kang et al., 2008; Zhang et al., 2010; Pankratz et al., 2005), or frailty models (Dey et al., 2022), which are more effective than PCs alone, but are computationally prohibiting.

2.3 Omitting covariates

The CPH model relies on the fundamental assumption that the effect of each covariate is proportional over time and relative to other covariates in the model. Although covariate effects can be approximately proportional in reality, apparent non-proportionality often results from model misspecification (Gail et al., 1984). Even in randomized trials where all known and unknown confounders are balanced between study arms, omitting a covariate can lead to bias in treatment effect estimates. This bias is particularly problematic in observational studies, where



neither the observed covariates nor unmeasured confounders are balanced across groups with different exposure levels (or genotypes). Consequently, the risk of omitted variable bias is significant, potentially leading to asymptotic biases. This results in systematically incorrect estimates, which can mislead conclusions about the relationships between variables. The severity of bias depends on the distribution of the omitted covariate, strength of its effect and censoring. Although one cannot directly adjust for unmeasured covariates, their potential impact can be assessed by sensitivity analyses (Lin et al., 2013), but that is hardly ever done in GWAS setting.

2.4 Other issues

In addition to the problems mentioned above, there are various other potential sources of bias that may affect the final conclusions, depending on the research question-we list them here for completeness, but ignore them in subsequent analysis.

When the outcome of interest is not death, but an incident condition, one should be aware of competing risks, such as death due to another cause. While treating competing events as censoring, however, the hazard ratios are still unbiased in general (Therneau and Grambsch, 2000), but care should be taken in absolute risk estimation tasks.

The assumption of proportionality of hazards has been discussed above in the context of omitted covariates, but the risk factors (including genetic variables) themselves may also have a time- or age-varying effect on hazard. However, this is again an issue that becomes important in absolute risk prediction, while for variant discovery one may accept that an average effect over time is estimated. Stensrud and Hernán (2020).

Recently it has been pointed out by several authors, that population-based biobanks are mostly non-random samples and therefore subject to selection bias (Schoeler et al., 2023; Lee and Han, 2022; Pirastu et al., 2021; van Alten et al., 2024). We agree that it is an important issue that should be taken into consideration in biobankbased studies, regardless of the type of variable (time to event or other) or method of analysis.

3 Ways to increase power and computational efficiency

3.1 Parent-offspring data

Joshi et al. (2016) have combined parent-offspring data to increase power of discovery. Biobank cohorts with short average follow-up time are underpowered for the analysis of participant lifespan data, due to the low number of outcome events (deaths). However, if family history at recruitment is collected and parental ages at death are known, they can be combined with subjects' genotype information.

If the age span of recruited subjects is sufficiently wide, a large proportion of them is likely to have parents who are either relatively old or already deceased. Therefore the use of parental data leads to lower censoring rates and higher power for genotype effect detection. As each allele in a SNP is inherited by offspring with



the probability of 50%, one can assume that using parental lifespan along with offspring genotype will result in estimates with magnitude of about half of the true effect size.

However, we have shown that the proportionality assumption in this case does not hold. We have derived equations for bias and show that the bias will increase, if the minor allele frequency and/or the effect size β increases. The derivation is explained in detail in the Supplementary Section 1.

3.2 Two-stage modeling via martingale residuals

To overcome computational challenges and leverage GWAS tools tailored for continuous phenotypes (Mbatchou et al., 2021;

Zhou et al., 2018; Loh et al., 2015; Jiang et al., 2019), we will examine performance of a two-step modelling approach proposed by Joshi et al. (2016) for a biobank setting. The idea of the method is fairly simple–instead of running a CPH model for every SNP, a single CPH model encompassing all the nongenetic and technical covariates is fitted in the first step. For that model, martingale residuals (MR) are obtained (Supplementary Section 2.3.2). As pointed out by Therneau et al. (1990), the association between MR and a covariate omitted from the linear predictor of the initial model yields estimates that align with the coefficients in the CPH model. Thus a test of a linear association between the MR and a genetic variant could potentially be used to detect an association between the variant and the outcome phenotype, reducing the association testing to a simple linear regression task.



Bias and power of various models across three timescales (TB, TR + A, TA), different censoring rates (CR), and MAFs fitted on a cohort with related individuals. The models include full covariate models with all subjects (CPH_F) and only independent subjects (CPH_{FNR}), models omitting the frailty term for all subjects (CPH) and only independent subjects (CPH_{RR}), and a CPH frailty model for all subjects (Frailty).

4 Results of the simulation study

Building on the sources of bias described in Section 2 and the strategies to improve power and computational efficiency discussed in Section 3, we now simulate different scenarios similar to real-life biobank data in order to determine if and how the above-mentioned methodological choices in the CPH model affect the results.

We will study the bias and power under various minor allele frequencies (MAF), effect sizes and censoring rates. Timescales are the following:

- timescale TB-time since birth;
- timescale TR + A-time since recruitment, age-adjusted;
- timescale TA-age as timescale (accounting for left-truncation).

To simulate time-to-event data, we generated one genotype, one additional covariate and survival times from a Weibull distribution. Survival time distribution was calibrated to match overall survival in the Estonian Biobank. We also assigned random birth and recruitment years to each individual. Individuals who would have died before recruitment were excluded to mimic left truncation. Right-censoring was imposed at different study cut-offs to vary censoring rates. Simulations were repeated 500 times on samples initially sized at 100,000, with final sizes varying due to exclusions. Full methodological details are provided in the Supplementary Section 2.1.

Our main aim is to compare effect size and significance of a SNP using the conventional CPH model and two-step MR approach. The impact on the working range of the approach is examined. The study aims to determine the effects of censoring and MAF on the performance of the two-step MR approach.

We will compare the models performances by:

- bias difference between real effect size and the estimated effect size;
- power probability of detecting a significant effect, when it is present;
- coverage probability that the true effect size lies in the confidence interval of the estimated effect size.



4.1 Effect of timescale choice on bias and power in CPH model

CPH with TA (Figure 1) is the only one that results in unbiased estimates for β_1 . In addition, this model exhibits the best coverage of the 95% confidence interval for true effect size.

CPH with TB results in the greatest bias, whereas the bias for TR + A case is very small. The bias for both TB and TR + A increases as effect sizes and censoring rates grow. Coverage of the true effect size for TB drops to zero for common variants (MAF = 0.4) already at $\beta_1 = 0.2$. The coverage for TR + A is not as good as for TA, although the differences are minor. Coverage can be seen to be better consistently when MAF is low.

For power we only present the plot, where MAF = 0.05 and CR = 90 (Figure 2) as the differences in the results are greatest here, other plots can be found in the Supplementary Figure S1.

CPH with TB results in the highest power to detect a significant association, whereas the power for TA is lowest no matter what the effect size. The differences in power for TB and TR + A can be up to 25%. CPH with TR + A and TA have very similar power regardless of the effect size.

As a conclusion we see that although the TB approach leads to potentially biased estimates of the true effect, it may be the preferred approach if the aim is to maximize power in a discovery study.

4.2 Utility of martingale residuals based approach in approximating CPH model estimates

As shown before, CPH with TB could be preferable in GWAS settings due to highest power to detect relatively small effect sizes.



Timmers et al. can be found here: https://doi.org/10.7554/eLife.39856.015.

Therefore, the simulation results on the performance of the MR approach are here presented only for TB (Figure 3), whereas the results for CPH with TR + A and TA are presented in the Supplementary Figures S3, S4.

A comparison of MR estimates with those from the standard CPH model-fitting algorithm and the actual effect sizes reveals that the two-step approach approximates CPH estimates quite well within typical GWAS working range. Compared to CPH estimates, the effect size estimates based on MR approach demonstrate less bias and higher relative accuracy in capturing the true parameter estimates. However, the relationship between bias and the true effect size exhibits a non-linear pattern, and the same nonlinearity holds for the coverage of the 95% CI. The observed nonlinearity requires further theoretical investigation. Similarly to the effect sizes, p-values obtained from the MR approach were more conservative than the ones obtained from the CPH model within typical GWAS setting. We observed in our simulation that when censoring rates were low, MR p-values approximated CPH p-values better, compared to settings with high censoring rates. Nevertheless, the power is rather similar to the CPH across different censoring rates and MAF for effect sizes within GWAS working range (see Supplementary Figure S2). Type I error is well controlled across all approaches, with no inflation of false positives observed under any scenario. Detailed estimates with 95% confidence intervals are provided in the Supplementary Table S3.

4.3 Relatedness and model misspecifation

To investigate issues regarding relatedness and model misspecification, we simulate datasets of siblings with three covariates to mimic realistic genotype, phenotype, and a shared family frailty, which is often unmeasured. The simulation details are provided in the Supplementary Section 2.2. We run analyses using three different timescales (TB, TR + A, TA), both ignoring relatedness (i.e., including relatives) and using only unrelated individuals. For each scenario, we compare models including all covariates to those omitting the frailty term. Additionally, we evaluate a CPH model with a frailty term for data including relatives, but ignoring frailty term as it would be in a realistic setting (Figure 4). Results indicate that ignoring relatedness does not significantly increase bias in effect size compared to the choice of timescale, regardless of censoring rate or MAF. However, omitting a covariate creates substantial bias and, in our setting, even changes the direction of the bias. The frailty CPH model shows the smallest

10.3389/fgene.2025.1534726

bias when using an age-adjusted timescale, with sensitivity to censoring rate and MAF—the smaller these two, the smaller the bias. Power analyses reveals that highest power for large censoring and small MAF occur with a birth-based timescale. Power is sensitive to censoring rate, MAF, ignoring relatedness, and omitting covariates, with the greatest decrease observed when both relatedness were ignored and covariates omitted. Notably, the coverage of the 95% CI decreases sharply when a covariate is omitted and relatives are included in the analysis, particularly when left-truncation is accounted for. This effect is more pronounced if the censoring rate decreases and MAF increases, due to the underestimation of variance (see Supplementary Figure S5).

4.3.1 Martingale residuals in relatedness and model misspecification

We investigate how the two-step martingale residual approach would work in a realistic setting by using a simulation in which we intentionally omit a covariate representing frailty. We are interested in whether the two-step martingale residual approach could approximate CPH estimates, specifically using age since birth as the timescale and under high censoring conditions. This investigation is conducted for both related and unrelated subjects. For this setup, the martingale residual approach leads to the smallest bias and highest coverage, although it has slightly less power than the standard CPH models (Figure 5). Therefore, for explanatory GWAS using age since birth as the timescale, the martingale residual-based approach appears to be a robust method for estimating hazard ratios, effectively handling related subjects and being computationally efficient.

5 Application to the Estonian Biobank data

The Estonian Biobank maintains a volunteer-based cohort of the Estonian adult population (aged ≥ 18 years) (Leitsalu et al., 2015; Milani et al., 2025). The sample size used in this analysis is 51 463, which represents approximately 5% of the Estonian adult population (participants recruited during the first period of recruitment in 2002–2011). In this sample, 65.6% of participants were female and the median age at recruitment was 43 (min = 18, max = 103) years. Median follow-up time with IQR was 13.1 (11.7; 13.9) years. The lifespan data of the participants is obtained via record linkages with the Estonian Causes of Death Register (latest linkage for the data used here was in the beginning of 2022). The mortality rate in the analysed sample was 13.2%.

Testing the top 11 SNPs and the polygenic risk score (GRS) for lifespan based on Timmers et al. (Timmers et al., 2019), we fit models with three choices for timescale: time since birth (TB), time since recruitment with age-adjustment (TR + A) and age as timescale (TA, accounting for left-truncation). For each timescale choice, we fit the model for the entire sample (ignoring the relatedness) and also for the sample where relatives with identity by descent greater than 20% were excluded (the remaining sample size: n = 38 223). In addition to age, all the models are also adjusted for sex. Thus as a result, 6 different models are fitted in total. Due to differences in sample size, the estimates from Timmers et al. have narrower confidence intervals than those from the Estonian Biobank.

Out of the 3 top SNPs presented in Figure 6, tested, only one (APOE) shows significant result in the Estonian Biobank. The results for all 11 SNPs can be found in Supplementary Table S4. As expected the GRS shows the greatest effect on mortality, whereas its effect size is almost identical regardless of whether the relatives are included or excluded. Largest effect size estimate is obtained when TB was used as the timescale–as pointed out before, the bias due to left-truncation is a likely cause for this difference. The other two timescale choices do not lead to visible differences in the effect sizes. The estimates of the effects of the LPA, APOE and CHRNA3/5 variants do not differ much, but excluding relatives has generally reduced the estimated effect sizes for APOE and CHRNA3/5, whereas the timescale choice does not really have any clear impact.

6 Discussion

For accurate estimation of population parameters, unbiasedness is the essential requirement for any statistical estimators. However, the task of estimation should be distinguished from the task of hypothesis testing. The present work has highlighted that in the context of GWAS for time to event outcomes, the estimators leading to the smallest bias are not necessarily the ones corresponding to most powerful tests for the hypothesis of no genotype-phenotype association.

Time-to-event phenotypes are challenging for GWAS, as the commonly used analysis tools, such as CPH modeling, require considerably more computational resources for implementation than algorithms for linear and logistic regression analysis. In addition, as the power depends not on the total sample size, but on the number of (disease or mortality) events observed, even a large biobank cohort may not be sufficiently powered for the discovery of biologically meaningful outcome-associated variants. Thus approaches that maximize power are especially welcome for time-to-event GWAS analyses, even if they come at the cost of some bias in parameter estimates.

The first finding of the present study is, that although careful adjustment for left-truncation is needed to achieve unbiasedness, it would considerably reduce power in a discovery GWAS compared to the approach that ignores it. For instance, a true hazard ratio of 1.05 (typical effect size of a common variant in GWAS) is likely to be overestimated by 2%–3%, whereas the power to detect an association may be increased by more than 1.5 times, when time since birth is used as timescale and left-truncation is ignored.

As under the null hypothesis of zero effect size the bias could not occur, ignoring left-truncation would not increase type I error probability. There is, however, one exception-the case where a genetic variant has been under selection. By "enriching" the risk sets with individuals recruited at later time points, one may in these cases create a situation where the allele frequencies in subjects with outcome events differ systematically from the allele frequencies in the risk sets. We recommend that this issue should be examined for variants identified as significant in a GWAS.

To simplify the computational algorithm of model-fitting, the use of a two-step procedure involving martingale residuals has been explored in the GWAS context. Martingale residuals were initially proposed as a diagnostic tool for a CPH model (mainly to identify appropriate covariate transformations), and to our knowledge, their

10.3389/fgene.2025.1534726

use in the actual effect estimation has not been explored in detail. We have shown that the two-step procedure provides valid estimates with no (or negligible) bias for the estimation of the relatively small effect sizes that are typical for GWAS findings. In addition, we have also shown that using the MR approach the power of the association discovery is not decreased compared to the corresponding CPH model.

Based on our simulations, ignoring relatedness does not significantly increase bias in effect size, whereas omitting key covariates introduces substantial bias. Additionally, the two-step martingale residual approach proved to be robust for estimating hazard ratios, efficiently handling related subjects with high coverage and slightly reduced power compared to standard CPH models.

In summary, our results support that for a time-to-event phenotype, a procedure where: 1) age is used as a time-scale and left-truncation is ignored and 2) a two-step procedure that obtains martingale residuals at the first step and runs a linear regressionbased GWAS as the second step is implemented, leads to better computational efficiency and better power for variant discovery than the procedure that fits a CPH model separately for each variant, whereas adjusting for other covariates.

Once the set of potentially associated variants is identified, we still recommend to validate the findings in both the discovery cohort(s) and also in a large independent cohort, using the CPH modeling approach that leads to unbiased estimates (thus, properly accounting for left-truncation). The latter is especially true, when the effects of polygenic risk scores (GRS) are estimated, as biases in these estimates are not acceptable when personalized risk prediction algorithms are derived. Also, to compute a GRS based on estimated regression coefficients from GWAS, one needs unbiased estimates for those coefficients.

Data availability statement

The data analyzed in this study were obtained from the Estonian Biobank. The data are not publicly available due to participant confidentiality and national regulations. However, the Estonian Biobank is open to researchers worldwide, and data access is granted through a well-established application process (https:// genomics.ut.ee/en/content/estonian-biobank).

Ethics statement

The activities of the Estonian Biobank are regulated by the Human Genes Research Act, which was adopted in 2000 specifically for the operations of the EstBB. Individual-level data analysis in the EstBB was carried out under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs), using data according to release application 6-7/GI/28055 from the Estonian Biobank. All participants of the Estonian Biobank have signed a written informed consent form.

Author contributions

AK: Writing - original draft, Writing - review and editing, Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization. MKo: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing original draft. Methodology, Writing review and editing. MKä: Writing - review and editing. MM: Methodology, Supervision, Writing - review and editing. KF: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing - original draft, Writing - review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. AK was supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 813533. KF, MKo and MKä received support from Estonian Research Council through grant PRG1197. MK received support from Estonian Research Council through grant PRG19911. This work was supported by the Estonian Research Council grant TARISTU24-TK19 and has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060011. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. ChatGPT was used to assist with text editing.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2025.1534726/ full#supplementary-material

10.3389/fgene.2025.1534726

References

Austin, P. C., Douglas, L. S., and Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation* 133 (6), 601–609. doi:10. 1161/CIRCULATIONAHA.115.017719

Bretagnolle, J., and Huber-Carol, C. (1988). Effects of omitting covariates in cox's model for survival data. *Scand. J. Statistics* 15 (2), 125–138. Available online at: http://www.jstor.org/stable/4616093

Cox, D. R. (1972). Regression models and life-tables. J. R. Stat. Soc. Ser. B Methodol. 34 (2), 187–202. doi:10.1111/j.2517-6161.1972.tb00899.x

Dey, R., Zhou, W., Kiiskinen, T., Havulinna, A., Elliott, A., Karjalainen, J., et al. (2022). Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nat. Commun.* 13 (1), 5437. doi:10.1038/s41467-022-32885-x

Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 71 (3), 431–444. doi:10.1093/biomet/71.3.431

Gogarten, S. M., Bhangale, T., Conomo, s M. P., Laurie, C. A., McHugh, C. P., Painter, I., et al. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* 28 (24), 3329–3331. doi:10. 1093/bioinformatics/bts610

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology* 13 (5). doi:10.1097/ EDE.0b013e3181c1ea43

Hughey, J. J., Rhoades, S. D., Fu, D. Y., Bastarache, L., Denny, J. C., and Chen, Q. (2019). Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. *BMC Genomics* 805 (20). doi:10. 1186/s12864-019-6192-1

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., et al. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* 51, 1749–1755. doi:10.1038/s41588-019-0530-8

Joshi, P. K., Fischer, K., Schraut, K. E., Campbell, H., Esko, T., and Wilson, J. F. (2016). Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan. *Nat. Commun.* 7 (1), 11174. doi:10.1038/ncomms11174

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178 (3), 1709–1723. doi:10.1534/genetics.107.080101

Korn, E. L., Graubard, B. I., and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *Am. J. Epidemiol.* 145 (1), 72–80. doi:10.1093/oxfordjournals.aje.a009034

Lagakos, S. W., and Schoenfeld, D. A. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* 40 (4), 1037–1048. doi:10. 2307/2531154

Lee, H., and Han, B. (2022). A theory-based practical solution to correct for sexdifferential participation bias. *Genome Biol.* 138 (23). doi:10.1186/s13059-022-02703-0

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M. L., Alavere, H., Snieder, H., et al. (2015). Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. *Int. J. Epidemiol.* 44 (4), 1137–1147. doi:10.1093/ije/dyt268

Lemieux Perreault, L. P., Legault, M. A., Asselin, G., and Dubé, M. P. (2016). genipe: an automated genome-wide imputation pipeline with automatic reporting and statistical tools. *Bioinformatics* 32 (23), 3661–3663. doi:10.1093/bioinformatics/btw487

Lin, N. X., Logan, S., and Henley, W. E. (2013). Bias and sensitivity analysis when estimating treatment effects from the cox model with omitted covariates. *Biometrics* 69 (4), 850–860. doi:10.1111/biom.12096

Loh, P. R., Tucker, G., Bulik-Sullivan, B., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* 47, 284–290. doi:10.1038/ng.3190

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* 53, 1097–1103. doi:10.1038/s41588-021-00870-7

Milani, L., Alver, M., Laur, S., Reisberg, S., Haller, T., Aasmets, O., et al. (2025). The Estonian Biobank's journey from biobanking to personalized medicine. *Nat. Commun.* 16 (3270), 3270. doi:10.1038/s41467-025-58465-3

Morgan, T. M., Lagakos, S. W., and Schoenfeld, D. A. (1986). Omitting covariates from the proportional hazards model. *Biometrics* 42 (4), 993–995. doi:10.2307/2530716

Novembre, J., and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* 40 (5), 646–649. doi:10.1038/ ng.139

Pankratz, V. S., de Andrade, M., and Therneau, T. M. (2005). Random-effects Cox proportional hazards model: general variance components methods for time-to-event data. *Genet. Epidemiol.* 28 (2), 97–109. doi:10.1002/gepi.20043

Pirastu, N., Cordioli, M., Nandakumar, P., Mignogna, G., Abdellaoui, A., Hollis, B., et al. (2021). Genetic analyses identify widespread sex-differential participation bias. *Nat. Genet.* 53, 663–671. doi:10.1038/s41588-021-00846-7

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi:10.1038/ng1847

R Core Team (2023). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/.

Rizvi, A. A., Karaesmen, E., Morgan, M., Preus, L., Wang, J., Sovic, M., et al. (2019). gwasurvivr: an R package for genome-wide survival analysis. *Bioinformatics* 35 (11), 1968–1970. doi:10.1093/bioinformatics/bty920

Schoeler, T., Speed, D., Porcu, E., Pirastu, N., Pingault, J.-B., and Kutalik, Z. (2023). Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat. Hum. Behav.* 7, 1216–1227. doi:10.1038/s41562-023-01579-9

Staley, J. R., Jones, E., Kaptoge, S., Butterworth, A. S., Sweeting, M. J., Wood, A. M., et al. (2017). A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *Eur. J. Hum. Genet.* 25 (7), 854–862. doi:10.1038/ejhg.2017.78

Stensrud, M. J., and Hernán, M. A. (2020). Why test for proportional hazards? *JAMA* 323 (14), 1401–1402. doi:10.1001/jama.2020.1267

Struthers, C. A., and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* 73 (2), 363–369. doi:10.1093/biomet/73.2.363

Syed, H., Jorgensen, A. L., and Morris, A. P. (2016). Evaluation of methodology for the analysis of 'time-to-event' data in pharmacogenomic genome-wide association studies. *Pharmacogenomics* 17 (8), 907–915. doi:10.2217/pgs.16.19

Syed, H., Jorgensen, A. L., and Morris, A. P. (2017). SurvivalGWAS_SV: software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes. *BMC Bioinforma*. 18 (265), 265. doi:10.1186/s12859-017-1683-z

Therneau, T. M., and Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. New York, NY: Springer.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* 77 (1), 147–160. doi:10.2307/2336057

Thiébaut, A. C., and Bénichou, J. (2004). Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics Med.* 23 (24), 3803–3820. doi:10.1002/sim.2098

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., and Risch, N. (2012). Estimating kinship in admixed populations. *Am. J. Hum. Genet.* 91 (1), 122–138. doi:10.1016/j.ajhg.2012.05.024

Timmers, P. R., Mounier, N., Lall, K., Fischer, K., Ning, Z., Feng, X., et al. (2019). Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *elife* 8, e39856. doi:10.7554/eLife.39856

van Alten, S., Domingue, B. W., Faul, J., Galama, T., and Marees, A. T. (2024). Reweighting UK Biobank corrects for pervasive selection bias due to volunteering. *Int. J. Epidemiol.* 53 (3), dyae054. doi:10.1093/ije/dyae054

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38 (2), 203–208. doi:10.1038/ng1702

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42 (4), 355–360. doi:10.1038/ng.546

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* 50, 1335–1341. doi:10.1038/s41588-018-0184-y