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

Merli Mändul*, Anastassia Kolde*, Meelis Käärik, Märt Möls, Krista Fischer

2023

Abstract

Survival analysis in clinical trials has been extensively researched, but its applicability to large omics-based biobanks requires further investigation. This study addresses two important issues in time-to-event data analysis in biobank settings. First, there is the need to pay attention to left-truncation of some outcomes that would prevent individuals joining the biobank, if the event of interest occurs before possible recruitment time. To avoid biases due to left-truncation, it has been suggested to use age as time scale in the analysis, whereas an individual is considered at risk only after recruitment to the biobank. Second, one needs to address the computational burden due to relative slowness of the conventional algorithm to maximize the partial likelihood function for the Cox model. This study examines the impact of the choice of timescale on the bias and power, while varying effect size and censoring rate in the range realistically seen in a typical biobank setting. Moreover, the study explores a computationally fast two-step martingale residual (MR) based approach for Cox modeling in high-dimensional omics data. The findings indicate that the choice of timescale has minimal impact on accuracy for small hazard ratios (HR), but for larger HRs, accounting for left-truncation is crucial to reduce bias. When maximizing power for association discovery, using participant age as the time scale (whereas

individual is considered to be at risk immediately after birth) yields the highest power. The two-step MR approach is recommended for genome-wide association studies due to acceptable precision and retained power for small effect sizes. However, for predictions and polygenic risk score calculations, effect sizes should be recalculated using the conventional Cox proportional hazards model while accounting for left-truncation. The conclusions are based on simulations and illustrated with survival data from the Estonian Biobank cohort.

1 Introduction

Survival analysis is a field of statistics with a long history. Whereas the origins of some methods (such as life tables) date back to centuries, a new era began in 1972, when Sir David Cox proposed the methodology of the semiparametric proportional hazards model, now often called as the Cox model [1]. Due to its relative simplicity and the fact that one does not need to specify a parametric failure time distribution, the Cox model has become a standard in the analysis of failure time data in the biomedical fields. Thus in the recent decades, when several large population-based biobanks have started to collect and analyse follow-up data, the Cox model would also be an obvious first choice for the statistical methodology.

While fitting the Cox model, it is crucial to assess the underlying assumptions, for the results to be valid. Many textbooks and statistics courses mainly focus on model-fitting in the context of (randomized) clinical trials. Thus the standard modelling option is to use time since study entry to an outcome event (often death) as the outcome variable of interest. The assumption of independent censoring is mostly fulfilled and due to a relatively short follow-up time, there are mostly no serious violations of the proportionality assumption of the hazard. While applying the methodology to different study designs, such as cohort studies, one needs to realize that often the standard set of modelling options needs to be modified to assure validity of the results. That is even more important while analysing the biobank cohorts.

First issue to be considered is the choice of the timescale. As shown already by Thiebaut et al. [2], in epidemiological cohort studies the appropriate timescale is age, and once used, no further age adjustment is needed. Still, this needs attention in the context of biobank data to understand not only the possible sources of bias, but also the implications on power, thus the bias/variance trade-off while making different timescale choices.

Power issues need to be discussed in a wider context as well, as the power in survival analysis is determined by the number of events of interest, whereas the total number of subjects plays a minor role. Therefore, even in large-scale biobanks it is sometimes difficult to achieve satisfactory power, as the follow-up time for the participants is not yet long enough to reach sufficient numbers of cases of incident diseases and mortality (especially for less common diseases or cause-specific mortality).

A third range of problems is related to the computational complexity of the Cox model. With genome-wide association studies (GWAS) sample sizes often exceeding 10 000 or even 100 000 individuals with more than ten millions of genetic markers, fitting Cox proportional hazards models using the conventional algorithms, is a time-consuming task even for the best computers. We will study a two-stage approach involving martingale residuals that dramatically reduces the computing time.

2 Methodological considerations in biobank data analysis

2.1 Choice of timescale in biobank data analysis

In most survival analysis textbooks the outcome variable (survival time) T is defined as time from a subject entering the study until the occurrence of the event of interest. The standard Cox proportional hazards model would then assume that all individuals have a common baseline hazard $h_0(t)$, where t denotes the time since study entry. The covariates (risk factors) are assumed to either

increase or decrease the hazard multiplicatively, so that a constant difference in one covariate corresponds to a constant hazard ratio throughout the follow-up time.

In biobank data, using age as timescale seems a more natural choice than time since recruitment, as the time of recruitment is not a meaningful timepoint regarding the outcome of interest. When assessing the effect of certain genetic variants, the question of interest is, whether individuals with the variant have in average higher or lower risk level than other individual of the same age. Also the natural baseline hazard can be seen to be a function of age, rather than a function of follow-up time. Still one could argue, that using time since study entry as timescale and age at recruitment as a covariate (thus accounting for the dependence of hazard on age), may be a valid alternative in some cases.

2.2 Age as timescale and left-truncation

Left-truncated survival data occurs when participants can only enter the study if they have already survived to certain time. Thus, while analysing biobank data, as highlighted by Thiebaut et al. [2], one should account for the fact that individuals who entered at the age of 70 years, say, do not form a representative sample of individuals born 70 years ago. This is rather a selective sample of those surviving at least 70 years. Therefore, based on the data of these particular individuals one cannot estimate the overall probability of surviving up to 80, for instance, but instead one might estimate the conditional probability of surviving to 80, given survival to 70.

In some cases left-truncation is ignored and age since birth is used as timescale in Cox models. We illustrate the source of bias with this kind of timescale usage by a simple example.

Suppose we have a binary risk factor, dividing individuals to "low" and "high" risk categories. Denote the hazard (of death or a disease) for low-risk individuals at age t, $h_0(t)$ and $\psi h_0(t)$ ($\psi \geq 1$) for high-risk individuals. Let p be the proportion of high-risk individuals at birth. Then it is easy to see, that in this

population, the ratio of high-risk to low-risk individuals at age t is a decreasing function of t:

$$\frac{p}{1-p} S_0(t)^{\psi-1},$$

where $S_0(t) = \exp(-\int_0^t h_0(u)du)$ is the survival function for low-risk individuals. To obtain unbiased estimates based on the cohort, the same ratio should hold in the risk sets within the cohort at any time t. If t is age at recruitment, then this would hold when the recruited individuals form a random sample of the population. However, if we include individuals that were recruited at age $t + \delta$ in the risk set for an event at age t, low-risk individuals will be over-represented in that risk set, leading to overestimation of the hazard ratio. Still, as no bias is anticipated under the null hypothesis ($\psi = 1$), the question remains, whether such approach (including in the risk set at time t all individuals known to be event-free at age t, regardless of their recruitment time) could be valid or even preferred for hypothesis testing – specifically, for identification of genetic variants associated with the time-to-event phenotype, due to better power than the unbiased alternatives.

2.3 Time since recruitment as timescale

Korn at al [3] have determined two sufficient conditions – when either one is satisfied, the age-adjusted Cox models with time since recruitment as timescale and Cox 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 these conditions are not fulfilled, Korn et al. suggests using Cox 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). However, Korn et al. did not provide a formal proof for the second condition, and Thiebaut et al. [2] found that mismodelling of age as an adjustment factor in a follow-up-dependent Cox model rather than as a timescale could also result in bias, in contrary to the Korn et al.'s reasoning. Other authors have shown that bias can be detected even when a variable independent from the variable of interest has been omitted from the Cox's regression model.

Thiebaut et al. also suggest using age as timescale rather than defining timescale based on observation time 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. Also, they point out that usually the number of subjects at risk does not vary monotonically with age, therefore the observation is conditional to the fact that the subject was free of disease at the age of recruitment.

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.

2.4 Two-stage modeling via martingale residuals

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, traditional Cox proportional hazards modeling has become computationally prohibitive and not easily scalable. To overcome computational challenges we will examine performance of a two-step modelling approach proposed by Joshi et al. (2016) [4] for biobank setting. The idea of the method is fairly simple – instead of running a Cox model for

every SNP, a single Cox model encompassing all the non-genetic and technical covariates is fitted in the first step. For that model, martingale residuals (MR) are obtained. As pointed out by Therneau et al. 1990 [5], the association between MR and a covariate omitted from the linear predictor of the initial model would indicate whether and in what functional form the covariate should be added to the 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 simple linear regression task.

In order to understand the bounds and limitations of the two-step approach for biobanks, we performed a simulation study. Our simulated dataset aims to correspond to Estonian biobank survival data.

3 Simulation study

3.1 Main objectives

We will simulate different scenarios similar to real life biobank data in order to determine if and how timescale choices in the Cox model affect the results. We will study the bias and power under various minor allele frequencies (MAF), effect sizes and censoring rates.

Knowing the ground truth, our main aim is to compare effect size and significance of a SNP using the conventional Cox model and two-step MR approach. The impact on the working range of the approach was examined. The study aimed to determine the effects of censoring and MAF on the performance of the two-step MR approach.

3.2 Data-generation algorithm

As the present study addresses two different and not directly related questions - the optimal timescale choice and appropriateness of the martingale residual approach, the data generation was performed separately, but in the similar man-

ner for the two parts of the study. Let's call these as part I and part II of the simulation study, respectively. We will highlight the minor differences in data generating processes below.

The parameters for the simulated distribution of the time to event variable are chosen to match the corresponding estimates from the overall survival distribution in the Estonian Biobank data.

The data-generation consists of two parts: part I for timescale and power analysis in Cox models and part II for martingale residuals based approach.

Step 1. We simulated the genotype variable $X_1 \sim Bin(2, MAF)$, where MAF represents the minor allele frequency of a SNP. Additionally, for part II we simulated one generic technical covariate $X_2 \sim N(0, 1)$.

Step 2. Survival time of the individuals was generated using Weibull distribution, $T \sim W(\lambda, \gamma)$. We chose the shape and scale parameters by the objective of generating survival times that closely approximate those observed in the Estonian Biobank. In part I the survival time was generated to be dependant only on the genotype. For part II the linear predictor was set to be a linear combination of simulated technical covariate and genotype. The effect size β_1 for genotype was varied mainly in the domain of GWAS effect sizes, to range from 0.01 to 0.91. The effect size of the technical covariate in part II was kept constant, $\beta_2 = 1$.

Step 3. We generated the year of birth for individuals in the population to be uniformly distributed between 1920 and 1990.

Step 4. We generated the year of recruitment from uniform distribution and identified the subset who were alive during the recruitment time and could potentially join the biobank. The range for recruitment years was varied in order to achieve different censoring rates.

The initial sample size for part I was $N = 100\ 000$, but for computational reasons the sample size for part II was smaller, $N = 10\ 000$. For each simulation the final sample size was varying due to omission of random number of people in step 4. These steps were repeated 500 times.

3.3 Models for simulated data

3.3.1 Part I. The effect of timescale choice on bias and power

In Cox PH models we looked at three different timescale scenarios:

1. models with age since birth as timescale, ie only accounting for right-censoring and ignoring the left-truncation (denoted **model A**);

2. age-adjusted models with time since recruitment as timescale (denoted **model B**);

3. models with age as timescale, ie accounting for both right-censoring and left-truncation (denoted **model C**).

Effect of timescale on bias are presented here for censoring rate 85%, whereas the results for censoring rates of 90% and 10% are presented in the Supplementary material. Power calculations are additionally done for censoring rates 95% and 80%.

3.3.2 Part II. Utility of martingale residuals based approach

To compare the conventional Cox model and two-step MR approach, the following models were fit:

1. The full model

$$h(t;x) = h_0(t)e^{\beta_1 X_1 + \beta_2 X_2}, \tag{1}$$

   where $h_0(t)$ is a baseline hazard function and X_1 represents genetic marker and X_2 a generic technical covariate using the conventional Cox partial likelihood approach for model-fitting.

2. Next we fit baseline Cox PH model where we account only for generic technical covariate $h(t;x) = h_0(t)e^{\beta_2 X_2}$, followed by calculation of the

martingale residuals

$$\hat{M}_i = \delta_i - \hat{H}_i(y_i|X_2 = x_{i2}),$$

where δ_i the event indicator for the i-th subject (1 - died, 0 - survived at the end of follow-up) and $\hat{H}_i(y_i|X_2 = x_{i2})$ is an estimate of the conditional expected value of δ_i for i-th subject for the total observation time y_i. As a final step in two-step approach, residuals, corrected for incidence, are regressed on the covariate of interest that was left out from the baseline Cox PH model

$$M^* = \alpha + \beta_1^* X_1, \tag{2}$$

where $M^* = M \cdot N / \sum_i (I(\delta_i = 1))$, $I(\delta_i = 1) = [\delta_i = 1]$, and N is sample size, whereas β_1^* should approximate β_1 from (1).

The performance of a two-step MR approach was evaluated under different censoring rates (93%, 82%, 78%, 60%, 34%, 11%, and 2%) and MAFs (0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5). The desired censoring rates were achieved by altering the length of the follow-up period.

4 Results

4.1 Part I: Effect of timescale choice on bias and power in Cox PH model

As seen from Table 1 model C was the only one that resulted in unbiased estimates for β_1. In addition, this model exhibited the best coverage of the 95% confidence interval for true effect size, as well as the most accurate estimates relative to the true values expressed as a $RMSE/HR$.

Model A gave the biggest bias, whereas model B resulted in a very small bias. The bias for both models A and B increased as hazard ratios grew. Coverage for model B was clearly better than for model A, where the coverage dropped to zero already when hazard ratio was set to be 1.2 and $MAF = 0.4$. The coverage

for model B was not as good as for model C, although the differences are worth mentioning only for bigger hazard ratios. Coverage can be seen to be better consistently when MAF is low, ie the rare variant scenarios.

The relative accuracy of estimates was better when $MAF = 0.4$, ie for common variants. The relative accuracy of the estimates was lowest for model A and remarkably better for model B, where it was only slightly worse than in model C.

		Model A	Model B	Model C
HR=1.05, MAF=0.05	**Bias**	0.01612	0.00431	0.00154
	RMSE/ HR	0.036	0.033	0.032
	Coverage of 95% CI	0.930	0.940	0.948
HR=1.05, MAF=0.4	**Bias**	0.01473	0.00254	-0.00054
	RMSE/ HR	0.020	0.014	0.014
	Coverage of 95% CI	0.840	0.936	0.952
HR=1.2, MAF=0.05	**Bias**	0.07117	0.01330	0.00141
	RMSE/ HR	0.078	0.039	0.036
	Coverage of 95% CI	0.564	0.938	0.948
HR=1.2, MAF=0.4	**Bias**	0.06972	0.01396	0.00061
	RMSE/ HR	0.069	0.022	0.017
	Coverage of 95% CI	0.020	0.852	0.950
HR=2, MAF=0.05	**Bias**	0.52056	0.03475	-0.00343
	RMSE/ HR	0.502	0.070	0.061
	Coverage of 95% CI	0.000	0.910	0.952
HR=2, MAF=0.4	**Bias**	0.51849	0.07342	0.00031
	RMSE/ HR	0.495	0.076	0.028
	Coverage of 95% CI	0.000	0.312	0.952

Table 1: Timescale choice simulations results with censoring rate 85%.*Model A- age since birth as timescale, model B - age-adjusted model with time since recruitment as timescale, model C - age as timescale.* Simulation results for censoring rates of 90% and 10% can be found in the Supplementary Table 1.

Regarding the power (Table 2), model A resulted in the highest power to detect a significant association with $HR = 1.05$ with every tested combination of censoring level and MAF, whereas the power for model C was lowest everywhere. With high censoring (95%) the power of model A was approximatley 1.6 times

higher than that of model C. For instance, of all SNPs with $MAF = 0.4$ and $HR = 1.05$, model C would detect 40.6%, model B 43% and model A close to two thirds.

Timescale	Censoring 95%		Censoring 90%		Censoring 85%		Censoring 80%	
	MAF= 0.05	MAF= 0.4	MAF= 0.05	MAF= 0.4	MAF= 0.05	MAF= 0.4	MAF= 0.05	MAF= 0.4
Model A	0.160	0.646	0.338	0.934	0.482	0.990	0.584	0.998
Model B	0.108	0.430	0.266	0.828	0.370	0.946	0.472	0.990
Model C	0.100	0.406	0.250	0.790	0.344	0.926	0.438	0.978

Table 2: Power for a model with $HR = 1.05$. *Model A- age since birth as timescale, model B - age-adjusted model with time since recruitment as timescale, model C - age as timescale*

4.2 Part II: Utility of martingale residuals based approach in approximating Cox PH model estimates

As shown before, model A could be preferable in GWAS settings due to highest power to detect relatively small effect sizes. Therefore, the simulation results on the performance of the MR approach are here presented only for model A, whereas the results for models B and C are presented in the Supplementary material (Figure 2 - 3).

A comparison of MR estimates of model A, B and C with those from the Cox model and actual ground truth HR-s reveals a clear trend that two-step approach approximates Cox PH estimates quite well within typical GWAS working range and depending on the censoring rate and choice of the time scale. Compared to Cox PH estimates, the HR estimates based on MR approach demonstrated less bias and higher relative accuracy in capturing the true parameter estimates.

When applying the MR approach for model A (Figure 1) as well as for models B and C (Supplementary Figure 2 - 3), the bias is even smaller as for the partial

likelihood estimates. Within GWAS working range, the bias appears to be negligible. Coverage of Cox PH estimates by MR based approach estimates' 95% CI decrease with the increase of HR, MAF, and decrease of the censoring rate. The relative accuracy of MR based estimates with respect to Cox PH estimates declined if HR increased and was highest within typical GWAS effect size range, regardless of censoring rate and MAF. With medium censoring rate MR based estimates for model A tend to be closer to the true HR rather than Cox PH estimates due to the systematic underestimation of the Cox PH parameters. Similarly, to the effect sizes, p-values obtained from the MR approach were more conservative than the ones obtained from the Cox PH model. We observed in our simulation that when censoring rates were low, MR p-values approximated Cox PH p-values better, compared to high censoring rates. Nevertheless, power is rather similar to Cox PH setting with respect to censoring rate and MAF within GWAS working range.

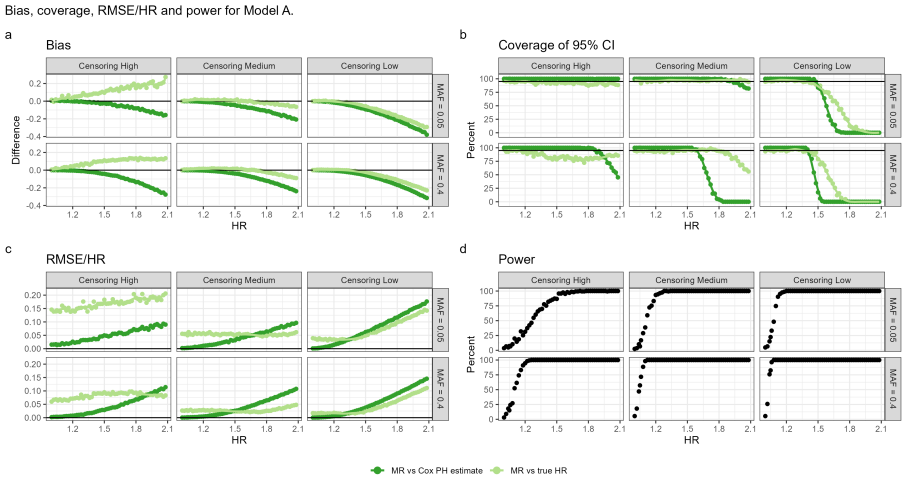

Figure 1: Performance metrics of Model A under different censoring rates and MAFs, encompassing the Bias, Coverage of 95% CI, RMSE/HR, and Power for HR. *Censoring High - 93%, Censoring Medium - 60%, Censoring Low - 11%*. Results for censoring rates 82%, 78%, 34% and 2% are presented in Supplementary Figure 1.

4.3 Application to the Estonian Biobank data

All models discussed before were also fitted to data from the Estonian Biobank. The Estonian Biobank is a volunteer-based cohort of the Estonian adult population (aged ≥ 18 years) [7]. The sample size used in this analysis was 51 463, which represents approximately 5% of the Estonian adult population. In the Estonian Biobank the lifespan of the recruited is followed via mortality registry and the latest linking to the registry was done in the beginning of 2022. The mortality among our sample was 13.2%.

We tested top 3 SNPs and a lifespan GRS that were found to be associated with lifespan according to Timmers et al. [6]. All models were adjusted to age and gender. 65.6% of the sample were female. The median age at recruitment was 43 $(\min = 18, \max = 103)$. Median follow-up time with IQR was $1.1(11.7; 13.9)$ years.

Out of the 3 top SNPs tested, only one SNP (APOE) showed significant result. This result stayed significant regardless of the timescale choice in the model. We also studied the effect of lifespan GRS on mortality. Similarly to APOE SNP this showed significant results regardless of the model choice.

Classical Cox models and Martingale residuals based approach resulted in almost identical estimates.

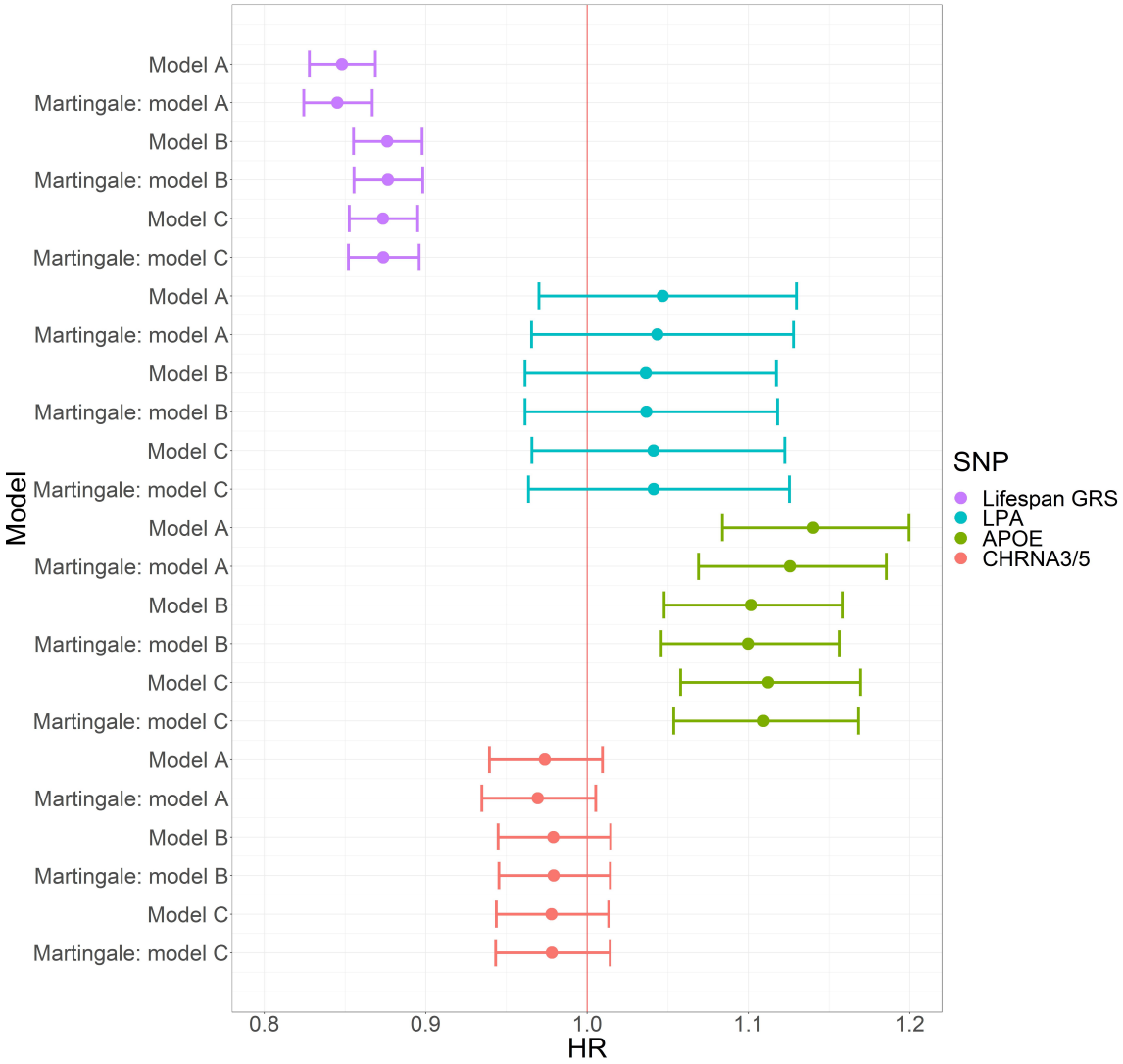

Figure 2: Comparison of hazard ratios for survival-associated SNPs and lifespan GRS using Classical Cox Models and respective MR approach in Estonian Biobank data. The figure shows HR estimates with 95% confidence intervals for Model A, B, and C on the x-axis, and the gene names on the y-axis. Gene names, rather than SNP names, are indicated on the y-axis. The numerical values of HR estimates, with 95% CI and p-values can be found in Supplementary Table 2.

5 Discussion

Unbiasedness has been the essential requirement for statistical estimators in virtually all statistics textbooks and articles so far. Indeed, one should definitely aim for it, if precise sample-based estimation of a population quantity is of primary interest. Still, one should distinguish the estimation task from a testing task – quite often these two are approached using the same methodology (such as regression modelling), but it does not necessarily have to be like that and maybe should not have to be like that. One example here is a GWAS, where the main task is to identify a subset of genomic variants that are associated with a phenotype, among millions of genotyped variants in a large population-based cohort.

Time-to-event phenotypes are challenging for GWAS, as their analysis requires survival analysis methods, such as Cox PH modeling, that require considerably more computational resources for implementation than algorithms for linear and logistic regression analysis. In addition, as the power of survival modelling 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 well-powered enough and methodological approaches that increase power, are needed.

The first finding of the present study is, that although careful adjustment for left-truncation is needed to achieve unbiasedness, it may not be necessary if avoidance of bias is not crucial, especially at the first discovery step of the analysis. The effect sizes of common genetic variants on common complex diseases or mortality tend to be rather small. We have seen that for hazard ratios (HR) in the magnitude of 1.05, the true HR is likely to be overestimated by 2-3%, whereas at the same time the power to detect an association may be increased by more than 1.5 times.

In general, the analysis that ignores left-truncation would not lead to increased type I error probability, as under the null hypothesis the bias could not occur. However, we can see one theoretical situation where type I error could still occur – namely, when a 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 observed events differ systematically from the allele frequencies in the risk sets.

To simplify the computational algorithm of model-fitting, the use of a two-step procedure involving martingale residuals has been shown to work well in the GWAS context. To our knowledge, the MR have been initially proposed as a diagnostic tool for fitted models, also to identify the potential relevance of omitted covariates – if suggested relevant, the covariates could then been added to the original Cox PH model. To our knowledge, the use of MR in the actual effect estimation has not been explored in detail. Here we have shown that the two-step procedure provides valid estimates with no (or negligible) bias, when the effect sizes are in the typical range for GWAS findings – thus, relatively small. In addition, we have also shown that the power of such a two-step procedure is not decreased compared to the corresponding Cox PH model.

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 regression-based GWAS as the second step is implemented, leads to better computational efficiency and better power for variant discovery than the procedure that fits a Cox PH 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 Cox PH modeling approach that leads to unbiased estimates (thus, properly accounting for left-truncation). The latter is especially true, when the effects of polygenic scores (PRS) are estimated, as biases in these estimates are not acceptable when personalized risk prediction algorithms are derived. Also, to compute a PRS based on estimated regression coefficients from GWAS, one needs unbiased estimates for those coefficients.

Acknowledgements

Anastassia Kolde was supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 813533. Krista Fischer received support from Estonian Research Council through grant PRG1197. This work was partially written at writing retreats and writing days organised by The University of Tartu Institute of Genomics.

References

[1] Cox, D.R., 1972. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), pp.187-202.

[2] Thiébaut, A.C. and Bénichou, J., 2004. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. Statistics in medicine, 23(24), pp.3803-3820.

[3] Kom, 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. American journal of epidemiology, 145(1), pp.72-80.

[4] 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. Nature communications, 7(1), p.11174.

[5] Therneau, T.M., Grambsch, P.M. and Fleming, T.R., 1990. Martingale-based residuals for survival models. Biometrika, 77(1), pp.147-160.

[6] Timmers, P.R., Mounier, N., Lall, K., Fischer, K., Ning, Z., Feng, X., Bretherick, A.D., Clark, D.W., Shen, X. and Esko, T., 2019. Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. elife, 8, p.e39856.

[7] Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L. and Fischer, K., 2015. Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. International journal of epidemiology, 44(4), pp.1137-1147.

Supplementary Material

Supplementary tables and figures

Table 1: Timescale choice simulations results with censoring rates 90% and 10%. *Model A- age since birth as timescale, model B - age-adjusted model with time since recruitment as timescale, model C - age as timescale*

	Timescale, censoring (%)	Model A, 90%	Model A, 10%	Model B, 90%	Model B, 10%	Model C, 90%	Model C, 10%
HR=1.05, MAF=0.05	**Bias**	0.01431	0.00429	0.00170	0.00301	-0.00061	0.00028
	RMSE/ HR	0.046	0.014	0.044	0.014	0.043	0.014
	Coverage for 95% CI	0.944	0.940	0.952	0.924	0.956	0.938
HR=1.05, MAF=0.4	**Bias**	0.01445	0.00410	0.00265	0.00314	0.00030	0.00032
	RMSE/ HR	0.025	0.007	0.021	0.007	0.021	0.006
	Coverage for 95% CI	0.874	0.898	0.942	0.912	0.944	0.948
HR=1.2, MAF=0.05	**Bias**	0.07389	0.01858	0.01340	0.01021	0.00289	0.00003
	RMSE/ HR	0.077	0.021	0.044	0.016	0.042	0.013
	Coverage for 95% CI	0.742	0.790	0.946	0.906	0.958	0.946
HR=1.2, MAF=0.4	**Bias**	0.07180	0.01846	0.01231	0.01145	0.00092	-0.00024
	RMSE/ HR	0.063	0.017	0.022	0.011	0.019	0.006
	Coverage for 95% CI	0.160	0.254	0.934	0.636	0.950	0.946
HR=2, MAF=0.05	**Bias**	0.57652	0.15015	0.05109	0.02194	0.00345	-0.00109
	RMSE/ HR	0.294	0.077	0.051	0.017	0.043	0.014
	Coverage for 95% CI	0.000	0.000	0.918	0.884	0.950	0.960
HR=2, MAF=0.4	**Bias**	0.60498	0.15290	0.08034	0.05435	-0.00120	-0.00051
	RMSE/HR	0.304	0.077	0.045	0.028	0.020	0.007
	Coverage for 95% CI	0.000	0.000	0.512	0.026	0.952	0.936

Table 2: Estonian Biobank analysis results. *Model A- age since birth as timescale, model B - age-adjusted model with time since recruitment as timescale, model C - age as timescale*

Timescale	SNP	Cox models		Martingale residuals based approach	
		HR (95 % CI)	p-value	HR (95 % CI)	pvalue
Model A	Lifespan GRS	0.8481 (0.8278; 0.8688)	$\mathbf{6.93 \cdot 10^{-41}}$	0.84530 (0.82437; 0.86677)	$\mathbf{2.45 \cdot 10^{-39}}$
Model B		0.8762 (0.8551; 0.8978)	$\mathbf{1.47 \cdot 10^{-26}}$	0.87660 (0.85554; 0.89818)	$\mathbf{2.73 \cdot 10^{-26}}$
Model C		0.8735 (0.8525; 0.8951)	$\mathbf{1.18 \cdot 10^{-27}}$	0.87370 (0.85208; 0.89588)	$\mathbf{4.88 \cdot 10^{-26}}$
Model A	c6p161010118 LPA (G/A, A AF=0.05)	1.0437 (0.9679; 1.1254)	0.2657	1.0436 (0.9656; 1.1279)	0.2817
Model B		1.0365 (0.9615; 0.1173)	0.3493	1.0368 (0.9615; 1.118)	0.3480
Model C		1.0412 (0.9657; 1.1225)	0.2926	1.0413 (0.9636; 1.1253)	0.3066
Model A	c19p45411941 APOE (C/T, T AF=0.13)	1.1299 (1.0745; 1.1882)	$< \mathbf{0.0001}$	1.1257 (1.069; 1.1855)	$< \mathbf{0.0001}$
Model B		1.1016 (1.0477; 1.1583)	**0.0002**	1.0997 (1.0459; 1.1564)	**0.0002**
Model C		1.1123 (1.0578; 1.1696)	$< \mathbf{0.0001}$	1.1095 (1.0536; 1.1684)	$< \mathbf{0.0001}$
Model A	c15p78817929 CHRNA3/5 (T/C, C AF=0.67)	0.9692 (0.9353; 1.0043)	0.0847	0.9694 (0.9348; 1.0053)	0.0939
Model B		0.9791 (0.9448; 1.0146)	0.2447	0.9793 (0.9453; 1.0144)	0.2443
Model C		0.9779 (0.9436; 1.0135)	0.2197	0.9782 (0.9433; 1.0143)	0.2327

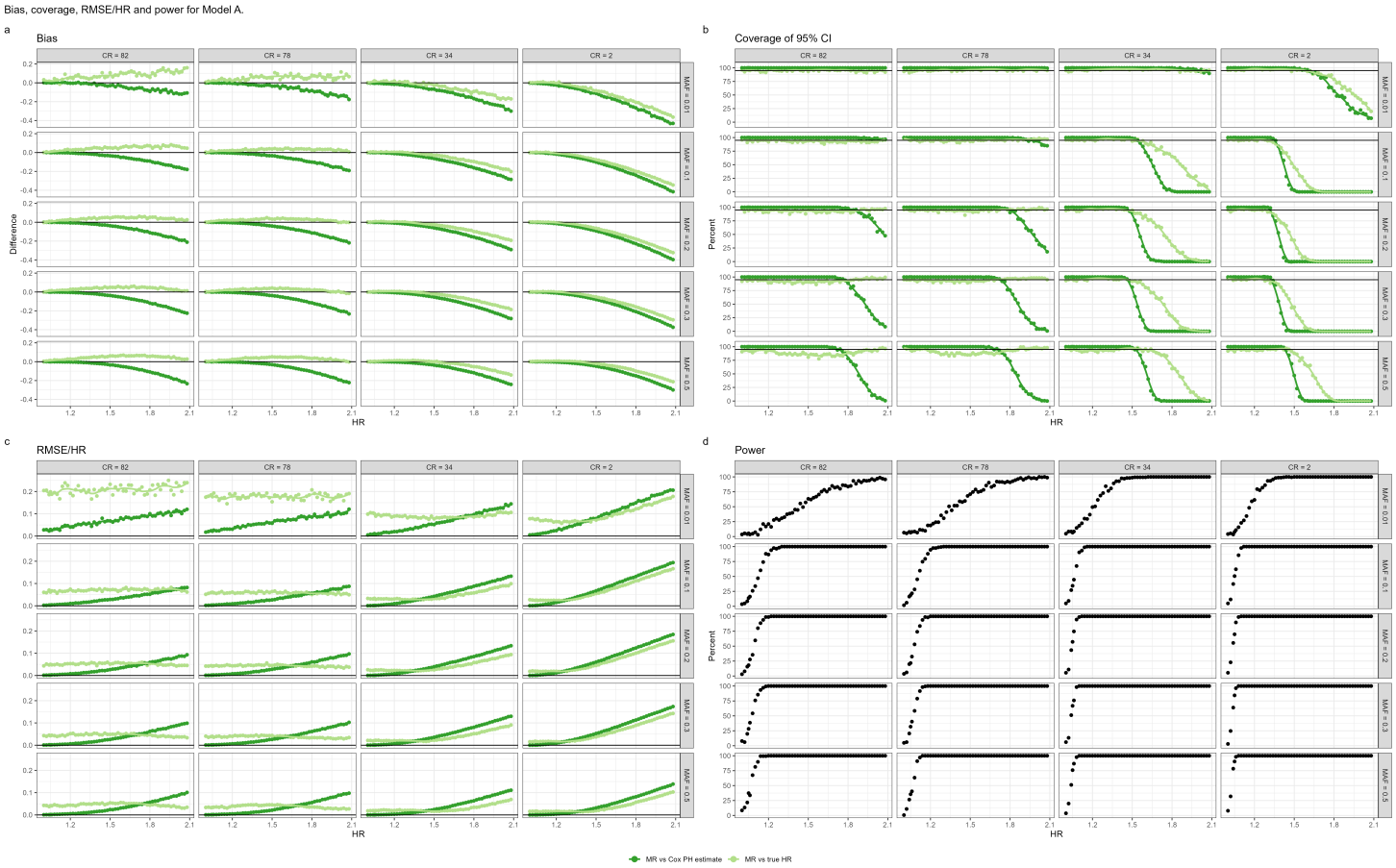

Figure 1: Performance metrics of Model A under different censoring rates and MAFs, encompassing the Bias, Coverage of 95% CI, RMSE/HR, and Power for HR. CR - censoring rate in %

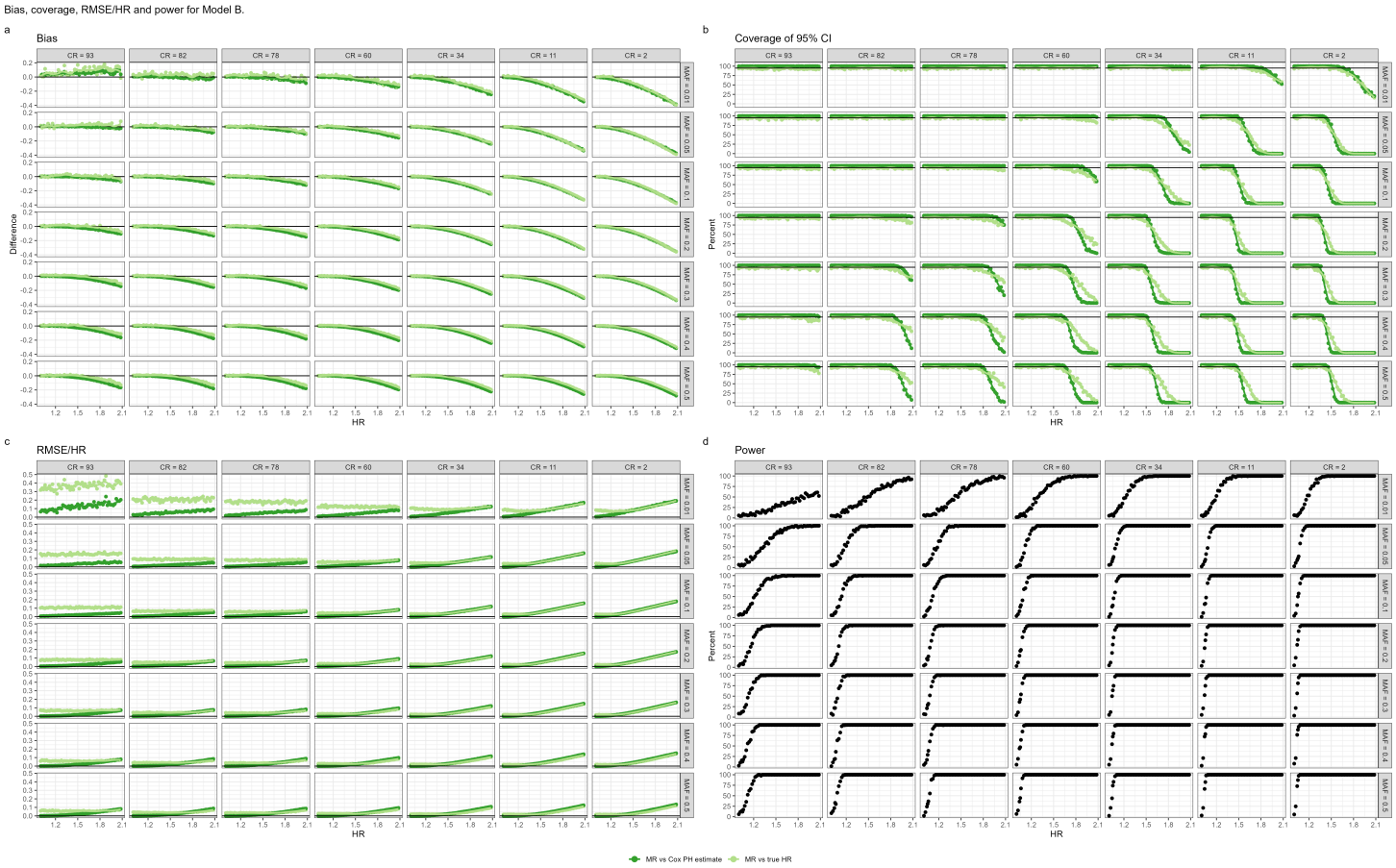

Figure 2: Performance metrics of Model B under different censoring rates and MAFs, encompassing the Bias, Coverage of 95% CI, RMSE/HR, and Power for HR. CR - censoring rate in %

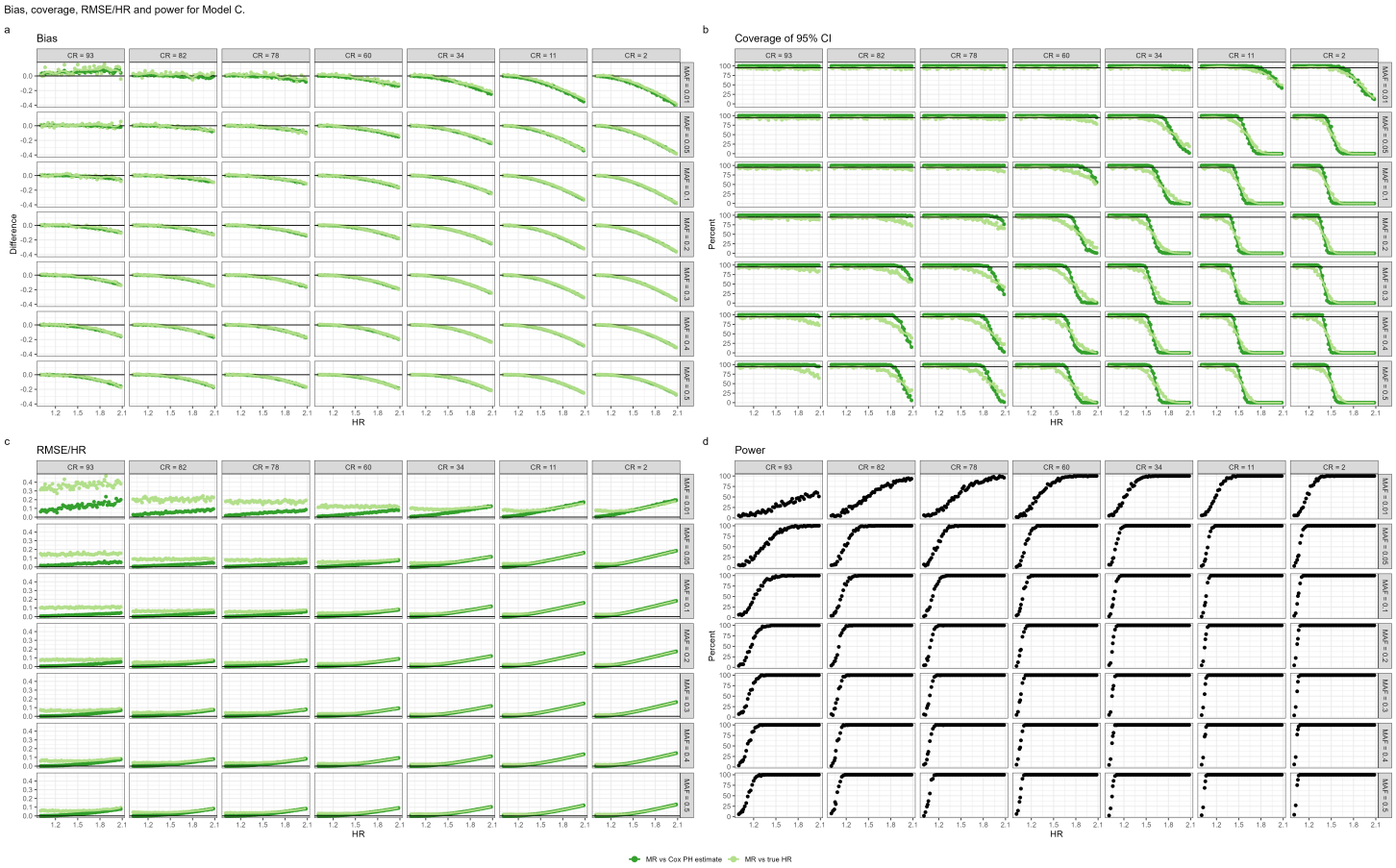

Figure 3: Performance metrics of Model C under different censoring rates and MAFs, encompassing the Bias, Coverage of 95% CI, RMSE/HR, and Power for HR. CR - censoring rate in %