Statistical Analysis of Self-Reported Health Conditions in Cohort Studies: Handling of Missing Onset Age

Sedigheh Mirazaei[*1], **José Miguel Martínez** [2], **Shizue Izumi** [3], **Motomi Mori** [1], **Gregory T. Armstrong** [4], **and Yutaka Yasui** [4]

[1] *Department of Biostatistics, St. Jude Children's Research Hospital, TN, USA.*

[2] *School of Public Health, University of Alberta, Edmonton, AB, CANADA.*

[3] *Faculty of Data Sciences, Shiga University, Hikone, Shiga, JAPAN.*

[4] *Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, TN, USA.*

Sedigheh.Mirzaei@stjude.org

S1 Childhood Cancer Survivor Study

The Childhood Cancer Survivor Study (CCSS), a retrospectively constructed cohort study with longitudinal, prospective follow-up of 25,735 5-year survivors of childhood cancer treated at 31 pediatric oncology institutions between 1970-1999 in the United States and Canada, provides a unique opportunity to study the late effects in pediatric cancer survivors. The study methods and design details have previously been described (Robison et al., 2009; Leisenring et al., 2009). The CCSS provides a wealth of information on the incidence and predictors of adverse health outcomes, in particular, survivor's self-reported chronic health conditions (CHCs) in a series of longitudinal surveys (`https://ccss.stjude.org/tools-documents/questionnaires.html`). In the CCSS, survivors are asked to recall the onset age of the CHC event if experienced. Some respondents report their ages at onset, while others leave it blank even when they indicated experiencing the CHC, which causes missing event onset age in the context of time-to-event analysis. The missingness occurs only among individuals who experienced the event, suggesting that the data is not missing data in the usual sense, and it is not certainly missing at random. Since the survivor reported the occurrence of the event at the time of the survey, the missing onset age is interval-censored, i.e., it is not just missing, but it is known to fall within the interval starting five years after diagnosis and extending to the last survey time, but its exact time is not known.

*To whom correspondence should be addressed.

S2 Missing onset age

A complete dataset without apparent bias cannot be created by simply eliminating the observations with missing data, which is tenable when the missingness is completely at random because the observations with a missing onset age had the event, and this missingness is not completely at random.

Various imputation methods, from non-statistically accurate single imputations to statistically valid multiple imputations, can address missing data issues, such as those involving missing onset ages, ultimately creating complete and analyzable datasets. Multiple imputation involves observed non-missing data to impute multiple values of missing data under certain assumptions on missing patterns and variable relationships. For the missing onset age problem, Taylor et al. (2002) proposed a non-parametric multiple-imputation method, further discussed by others (e.g., Gurney et al. (2003); Zhao et al. (2014).) Their multiple-imputation approach is effective under the specific assumptions of missing patterns and relationships among the variables used for imputation. As with any multiple imputation, their approach involves repeating the association analysis of interest for each imputed dataset and pooling multiple results using a special formula accounting for imputation-induced variability. Despite its effectiveness under assumptions made by researchers, the computational intensity and challenges in evaluating the tenability of assumptions limit its applicability in epidemiological research. Since the missing onset age is not at random, the existing statistical software's packages/procedures can not be used directly to address the missing onset age.

S3 Likelihood and method

The proportional hazards regression model (Cox, 1972) introduced flexibility in the analysis of time-to-event data, with its gain in popularity being attributed to its interpretability and ability to model right-censored data. Developing techniques that allow for the analysis of interval-censored data under semiparametric variants of this model can be challenging. These difficulties are encountered because of the underlying structure of interval-censored data, i.e., the event times of interest are never observed. In particular, data of this form typically consist of left-, right-, and interval-censored observations corresponding to the situation in which the event times occur before the first, after the last, or between two observation times, respectively. Interval-censored data is ubiquitous among social, behavioral, epidemiological, and medical studies Sun (2006), and therefore, modeling techniques that allow for the valid analysis of interval-censored data need to be developed, along with the necessary statistical software to carry out these analyses. The regression analysis of interval-censored data under the PH model is a well-studied problem. This problem was first addressed by Finkelstein (1986), who proposed a method of jointly estimating the regression parameters and the baseline hazard function using a Newton-Raphson-based algorithm. Subsequent developments in the interval-censored regression model include works by Groeneboom and Wellner (1992); Satten (1996); Goggins et al. (1998); Cai and Betensky (2003); Li and Ma (2013); Shao et al. (2014) among others. The Cox PH model simplifies the handling of right-censored data because it directly accommodates right-censored observations through the concept of risk sets, avoids the need to specify or estimate the baseline hazard function and benefits from broad support and availability in statistical software. The model and likelihood in these models are as follows.

Let $F(.|x)$ denote the cumulative distribution function (CDF) of the time-to-event of interest given the covariate vector $\boldsymbol{X}$. Under the PH model, the time-to-event distribution for individuals with

covariates $\boldsymbol{X_i}$ is given by , where $\boldsymbol{X_i} = (x_{i1}, \ldots, x_{ip})'$ is a $p \times 1$ vector of time-independent covariates, $\beta = (\beta_1, \ldots, \beta_p)'$ is the corresponding vector of regression parameters, and $\Lambda_0(.)$ is the cumulative baseline hazard function. It is assumed throughout that, conditional on the covariates, the time-to-event is independent of the observational process. This assumption is common in the survival literature; see, e.g., Sun (2006) among others. Under this assumption, the likelihood given the observed data $\mathcal{A} = \{(L_i, R_i|x_i)\}$, is

$$L_{obs} = \prod_{i=1}^{n} \{F(R_i^+|\boldsymbol{x_i}) - F(L_i^-|\boldsymbol{x_i})\}$$

where n is the sample size, L_i and R_i denote the left and right bounds of the observed interval for the ith individual, respectively, with $L_i < R_i$. Note, $L_i = 0 \quad (R_i = \infty)$ indicates that the ith individual's event time is left (right) censored and $L_i = R_i$ when the exact onset time is observed, Distinguishing between the three types of censoring, one can rewrite the observed data likelihood in the following form

$$L_{obs} = \prod_{i=1}^{n} \{F(R_i^+|\boldsymbol{x_i})\}^{\delta_{i1}} \{F(R_i^+|\boldsymbol{x_i}) - F(L_i^-|\boldsymbol{x_i})\}^{\delta_{i2}} \{1 - F(L_i^-|\boldsymbol{x_i})\}^{\delta_{i3}} \tag{1}$$

where δ_{i1}, δ_{i2}, and δ_{i3} are censoring indicators for the ith individual denoting left-, interval-, and right-censoring, respectively, subject to the constraint $\delta_{i1} + \delta_{i2} + \delta_{i3} = 1$, when $F(L_i^-)$ and $dF(R_i^+)$ are defined as $F(u^-) = \lim_{h \to 0} Pr(U < u - h); \quad F(u^+) = \lim_{h \to 0} Pr(U < u + h)$. Note that for the exact onset time (i.e., $L_i = R_i$), we define the probability mass at the exact onset time as $p_i = F(R_i^+) - F(L_i^-)$.

Under the assumption of proportional hazards, the probability of the individual i, with the covariate vector $\boldsymbol{X_i}$ and having the event after time t, is

$$\bar{\mathrm{F}}_i(t|\boldsymbol{X_i}) = [\bar{\mathrm{F}}_0(t)]^{\exp(\boldsymbol{\beta^T X_i})}, \tag{2}$$

where $\bar{\mathrm{F}}_0(t) = \Pr(T > t|\boldsymbol{X_i} = \boldsymbol{0})$ is the baseline survival distribution and $\boldsymbol{\beta^T}$ is the vector of parameters corresponding to the covariate vector under the Cox regression model.

Optimizing likelihood (1) after substituting (2) is computationally intensive or methodologically complex in practical applications. Wang et al. (2016) proposed a method for analyzing interval-censored data under the proportional hazards model, considering the following steps:

Step (1) The baseline cumulative hazard function, $\Lambda_0(.)$, was modeled using I-splines as proposed by Ramsay (1988).

Step (2) To employ an EM algorithm for finding the maximum likelihood estimates of the parameters, a two-stage data augmentation was used. This involved latent Poisson random variables, leveraging the relationship between the proportional hazards model and a nonhomogeneous Poisson process.

Step (3) An EM algorithm was used to optimize the augmented likelihood corresponding to the second stage of data augmentation.

Although not published in the Journal of Statistical Software, Wang et al. (2016) provided the `ICsurv` package in R for implementing their method in semiparametric regression analysis of

interval-censored data under the proportional hazards model. However, the selection of knots remains an open question for this package. Subsequently, Anderson-Bergman (2017) introduced a companion R package, `icenReg`, to facilitate fitting regression models for interval-censored data. In this study, to estimate the effects of risk factors for each chronic health condition, we utilized the '`ic_sp`' function from the `icenReg` package.

S4 Simulation Setup and Method

To assess the performance of methods described in Section 3, we conducted a comparative analysis of the estimated coefficients for the risk factors through Monte Carlo simulation For clarity, we introduce some notations used in the simulations.

Notation	Description
β	Regression coefficient (here we consider coefficients for two risk factors β_1, β_2)
$j = 1, 2, ..., 500$	Indexes the repetitions of the simulation
n	Sample size of a simulated dataset (here, $n = 300$, 1000, and 5000)
$F_0(t)$	Baseline distribution function
$\hat{\beta}$	The estimator of β
$\bar{\beta}$	The mean of β across repetitions
Bias	Difference between the true and estimated values of the regression coefficient $\text{bias}(\hat{\beta}) = (\hat{\beta} - \beta)$
Stdev	Standard deviation of the estimated regression coefficient (e.g. for β_1, it is defined as $\sqrt{Var(\hat{\beta_1})} = \sqrt{\frac{\sum_{j=1}^{500}(\hat{\beta_{j1}} - \bar{\beta_1})}{500}}$)
MSE	The mean square error of the estimated regression coefficient (e.g. for β_1, MSE (β_1) = var $(\hat{\beta_1})$+ (bias $(\hat{\beta_1}))^2$)

S4.1 Aims

This simulation study compares the performance of different methods introduced in Section 3 by comparing the bias, standard deviation, and MSE. Irrespective of the sample size, for the best method, bias, Stdev, and MSE should be as small as possible for the estimated regression coefficient β.

The assessment considered estimates from the 'observation-deletion,' 'event-deletion,' 'Simple replacement,' and the interval-censored regression ('Interval-censored'). Note that the '`coxph`' function in R or any Cox regression implementation in standard statistical software can be used for the first three approaches. The 'Interval-censored' approach used the '`ic_sp`' function in `icenReg` package of R. The '`ic_sp`' function fits a semi-parametric model for interval-censored data. We chose to fit a Cox proportional model. The covariance matrix for the regression coefficients is estimated via bootstrap; we used 50 bootstrap iterations. Note that this function has an option of parallel processing to take advantage of multiple cores for large datasets.

Based on the data described in Section 2, for the proportion of those who had the event, we consider two cases of **a) 10**% and **b) 30**%, using Binomial distributions with success probabilities 0.1 and 0.3. Among those who had the event, we consider three scenarios that could arise in our data regarding the missingness of onset age. **1) Equal Reporting/Non-reporting:** Among those who had the event, roughly 50% report the onset age, and approximately 50% do not report it.

2) Reporting Dominance: Among those who had the event, about 80% report their onset age, while approximately 20% do not report it. **3) Non-Reporting Dominance:** Among those who had the event, around 20% report the onset age, while the majority (80%) do not report it.

S4.2 Data-generation

In this simulation, we generate samples of time-to-event from a proportional hazards model with a survival function $S(t) = S_0(t)^{\exp\{\boldsymbol{\beta}' \boldsymbol{Z}\}}$, where the baseline distribution function $F_0(t) = 1 - S_0(t)$ is Weibull with shape and scale parameters $k = 10$ and $\lambda = 20$, respectively. This distribution has a median of about 20, aligned with a median time in years to experience some CHCs among childhood cancer survivors. We consider two independent covariates associated with the hazard rates of the CHC event: a binary variable, taking values 1 and 0 with probabilities 0.20 and 0.80, and a continuous variable with a uniform distribution over $[-5, 5]$. We choose the true vector of regression coefficients for the two covariates as $\beta = (\beta_1, \beta_2) = (1.5, 0.2)$. The 'time of survey/interview' is generated from the discrete uniform distribution over the set of integers $\{1, 2, \ldots, 35\}$. These choices align with the follow-up time in the CCSS data. Note that we consider the possible interval from the starting time to the survey time for the 'Interval-censored' approach.

S5 Simulation Results

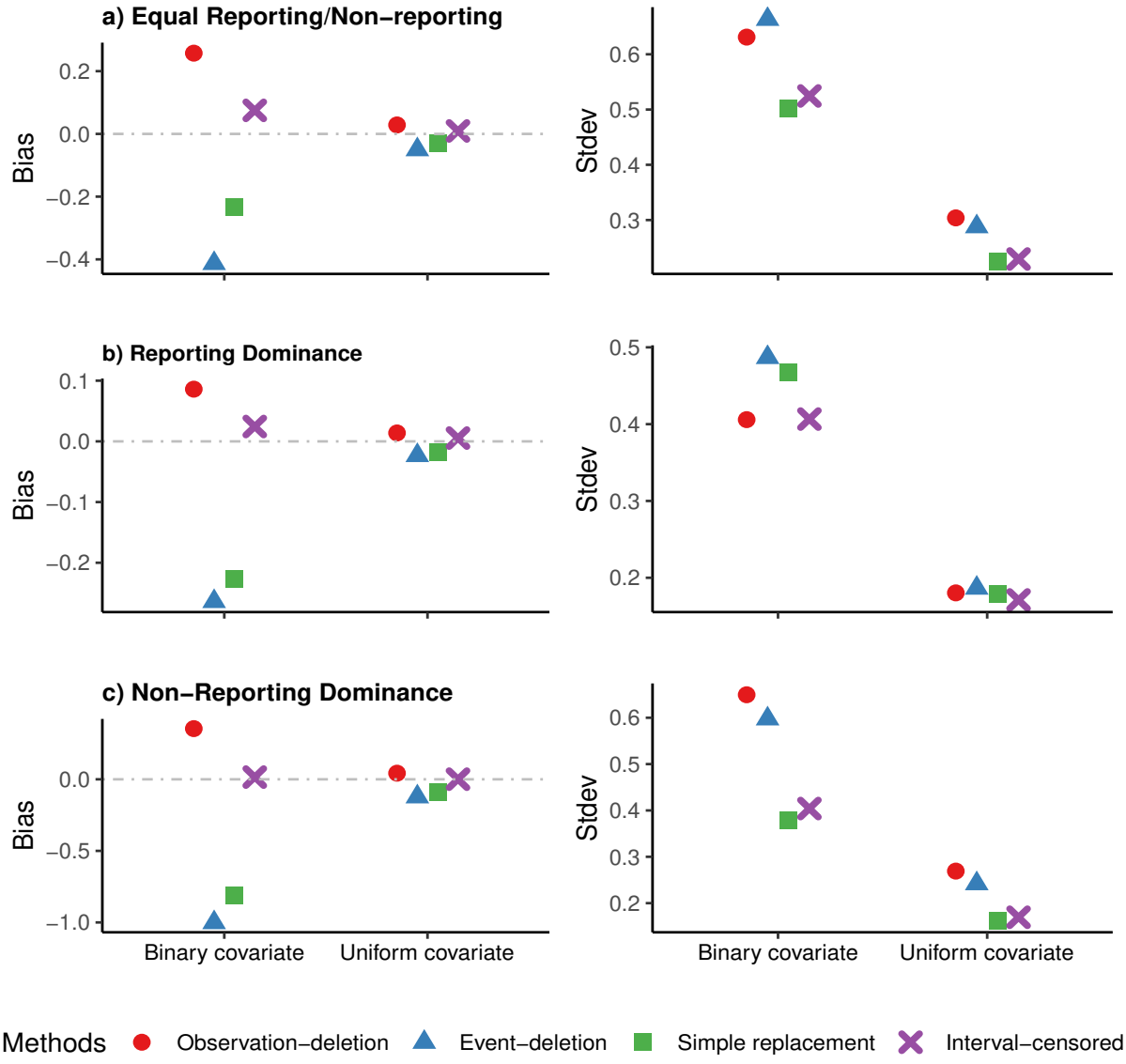

Figure S1: Plots of observed bias in the left column, and standard deviation (Stdev) in the right column of the estimated risk factor's coefficient using the Cox regression models 'observation-deletion,' 'event-deletion,' 'simple replacement,' and 'Interval-censored,' for a) 'Equal Reporting/Non-reporting' (the top panel), b) 'Reporting Dominance' (the middle panel), and c) 'Non-Reporting Dominance' (the bottom panel), for n=300 in 500 simulations when about 10% of individuals have experienced the event by the time of survey.

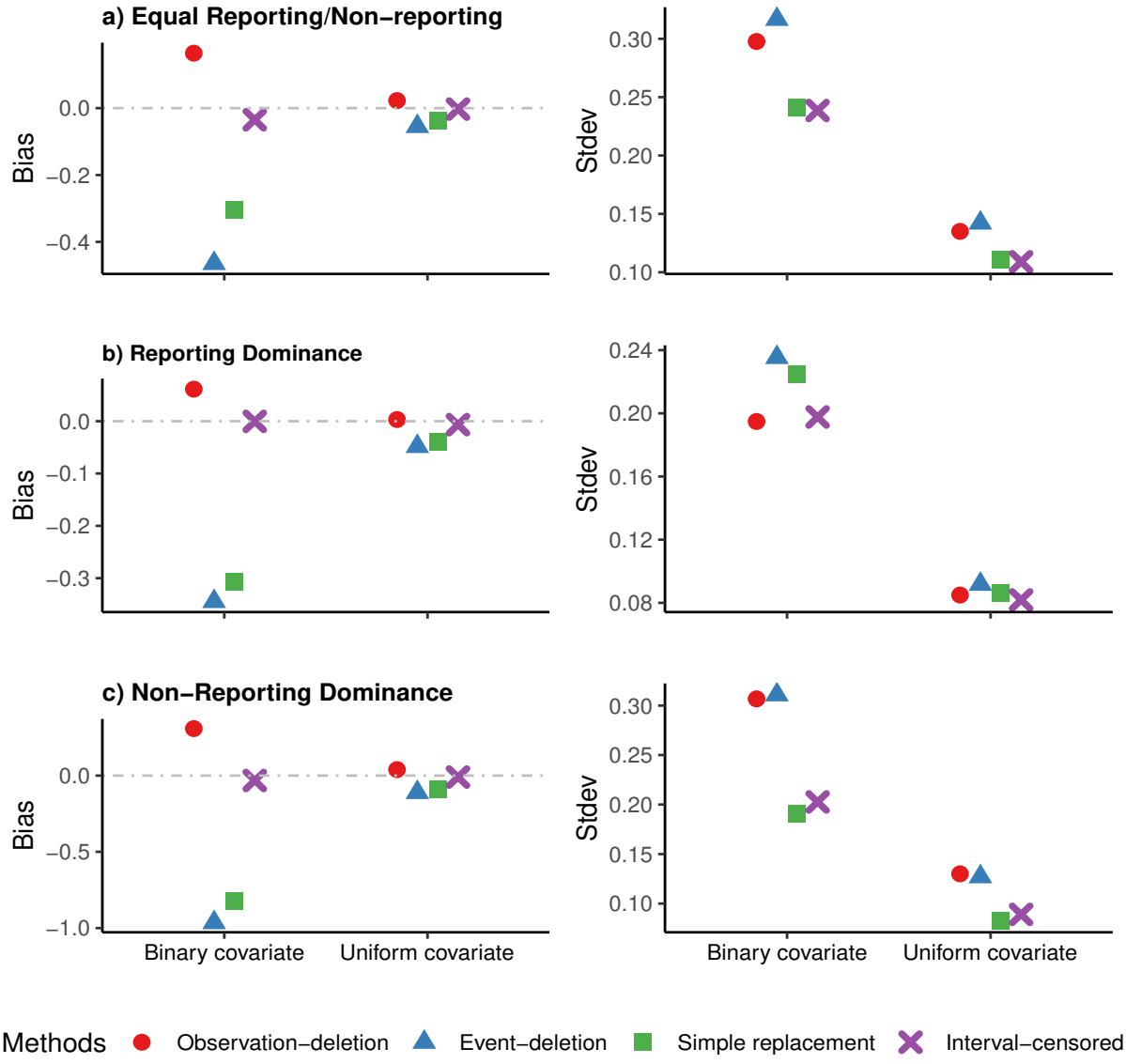

Figure S2: Plots of observed bias in the left column, and standard deviation (Stdev) in the right column of the estimated risk factor's coefficient using the Cox regression models 'observation-deletion,' 'event-deletion,' 'simple replacement,' and 'Interval-censored,' for a) 'Equal Reporting/Non-reporting' (the top panel), b) 'Reporting Dominance' (the middle panel), and c) 'Non-Reporting Dominance' (the bottom panel), for n=1000 in 500 simulations when about 10% of individuals have experienced the event by the time of survey.

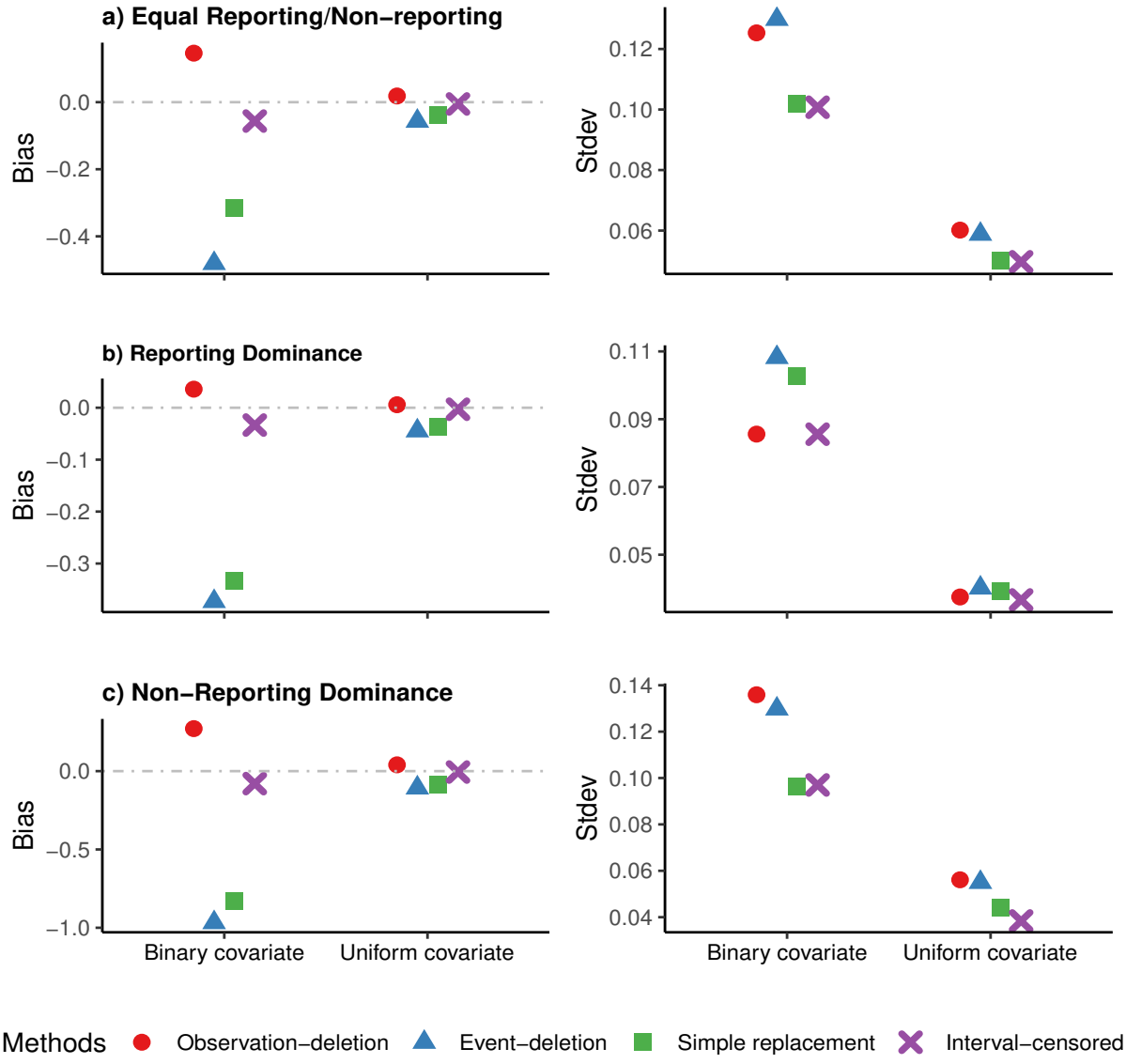

Figure S3: Plots of observed bias in the left column, and standard deviation (Stdev) in the right column of the estimated risk factor's coefficient using the Cox regression models 'observation-deletion,' 'event-deletion,' 'simple replacement,' and 'Interval-censored,' for a) 'Equal Reporting/Non-reporting' (the top panel), b) 'Reporting Dominance' (the middle panel), and c) 'Non-Reporting Dominance' (the bottom panel), for n=300 in 500 simulations when about 30% of individuals have experienced the event by the time of survey.

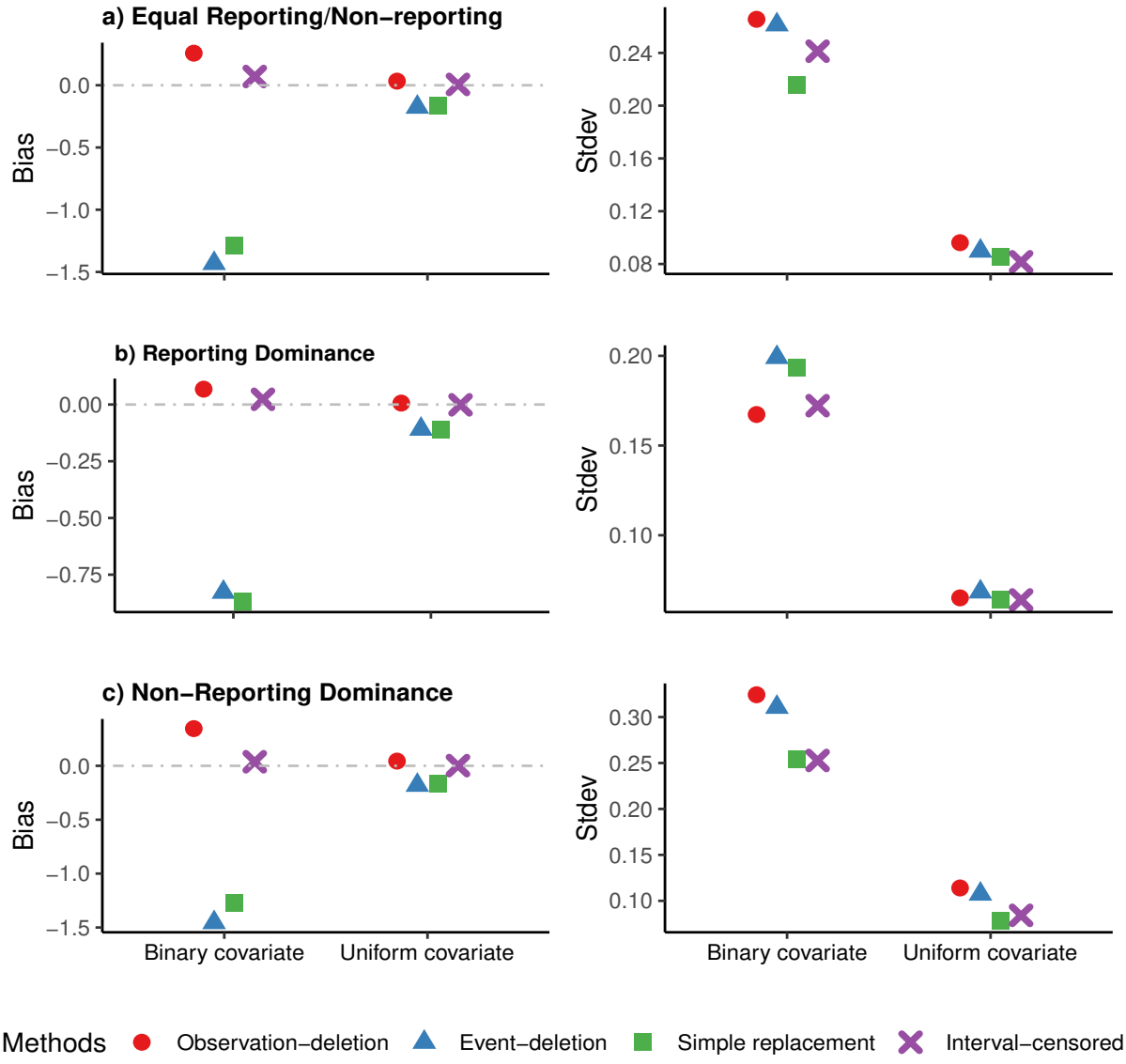

Figure S4: Plots of observed bias in the left column, and standard deviation (Stdev) in the right column of the estimated risk factor's coefficient using the Cox regression models 'observation-deletion', 'event-deletion', 'simple replacement', and 'Interval-censored', for a) 'Equal Reporting/Non-reporting' (the top panel), b) 'Reporting Dominance' (the middle panel), and c) 'Non-Reporting Dominance' (the bottom panel), for n=1000 in 500 simulations when about 30% of individuals have experienced the event by the time of survey.

Table S1: Bias, Stdev, MSE, and mean of standard error (SE) of the estimated coefficients for $n = 300$ and three scenarios in case (a)when about 10% of individuals have experienced the event by the time of survey.

Scenario	Method	Covariates	True value	Bias	Stdev	MSE	Mean SE
Equal Reporting/Non-reporting	Observation-deletion	Binary covariate	1.5	0.257	0.631	0.464	0.605
		Uniform covariate	0.2	0.029	0.304	0.093	0.271
	Event-deletion	Binary covariate	1.5	–0.413	0.663	0.610	0.561
		Uniform covariate	0.2	–0.050	0.288	0.085	0.262
	Simple replacement	Binary covariate	1.5	–0.234	0.502	0.307	0.444
		Uniform covariate	0.2	–0.030	0.224	0.051	0.212
	Interval-censored	Binary covariate	1.5	0.075	0.524	0.280	0.765
		Uniform covariate	0.2	0.010	0.229	0.053	0.253
Reporting Dominance	Observation-deletion	Binary covariate	1.5	0.086	0.406	0.172	0.381
		Uniform covariate	0.2	0.014	0.180	0.033	0.163
	Event-deletion	Binary covariate	1.5	–0.264	0.487	0.306	0.363
		Uniform covariate	0.2	–0.023	0.187	0.035	0.162
	Simple replacement	Binary covariate	1.5	–0.226	0.467	0.269	0.351
		Uniform covariate	0.2	–0.017	0.178	0.032	0.157
	Interval-censored	Binary covariate	1.5	0.024	0.407	0.166	0.420
		Uniform covariate	0.2	0.005	0.171	0.029	0.174
Non-Reporting Dominance	Delete no recall	Binary covariate	1.5	0.354	0.650	0.546	0.622
		Uniform covariate	0.2	0.043	0.269	0.074	0.252
	Event-deletion	Binary covariate	1.5	–1.001	0.598	1.359	0.537
		Uniform covariate	0.2	–0.122	0.242	0.073	0.233
	Simple replacement	Binary covariate	1.5	–0.810	0.379	0.800	0.334
		Uniform covariate	0.2	–0.092	0.162	0.034	0.155
	Interval-censored	Binary covariate	1.5	0.015	0.404	0.164	0.415
		Uniform covariate	0.2	0.001	0.170	0.029	0.183

Table S2: Bias, Stdev, MSE, and mean of standard error (SE) of the estimated coefficients for $n = 1000$ and three scenarios in case (a)when about 10% of individuals have experienced the event by the time of survey.

Scenario	Method	Covariates	True value	Bias	Stdev	MSE	Mean SE
Equal Reporting/Non-reporting	Observation-deletion	Binary covariate	1.5	0.164	0.298	0.116	0.299
		Uniform covariate	0.2	0.023	0.135	0.019	0.138
	Event-deletion	Binary covariate	1.5	–0.464	0.317	0.316	0.284
		Uniform covariate	0.2	–0.055	0.142	0.023	0.136
	Simple replacement	Binary covariate	1.5	–0.305	0.241	0.151	0.226
		Uniform covariate	0.2	–0.037	0.111	0.014	0.110
	Interval-censored	Binary covariate	1.5	–0.035	0.239	0.058	0.254
		Uniform covariate	0.2	–0.003	0.109	0.012	0.115
Reporting Dominance	Observation-deletion	Binary covariate	1.5	0.061	0.195	0.042	0.197
		Uniform covariate	0.2	0.003	0.085	0.007	0.085
	Event-deletion	Binary covariate	1.5	–0.345	0.235	0.174	0.186
		Uniform covariate	0.2	–0.048	0.092	0.011	0.084
	Simple replacement	Binary covariate	1.5	–0.307	0.225	0.144	0.179
		Uniform covariate	0.2	–0.040	0.086	0.009	0.081
	Interval-censored	Binary covariate	1.5	0.0003	0.198	0.039	0.199
		Uniform covariate	0.2	–0.007	0.082	0.007	0.083
Non-Reporting Dominance	Observation-deletion	Binary covariate	1.5	0.308	0.307	0.189	0.306
		Uniform covariate	0.2	0.039	0.130	0.018	0.128
	Event-deletion	Binary covariate	1.5	–0.964	0.311	1.026	0.277
		Uniform covariate	0.2	–0.111	0.127	0.028	0.124
	Simple replacement	Binary covariate	1.5	–0.822	0.191	0.711	0.175
		Uniform covariate	0.2	–0.089	0.082	0.015	0.081
	Interval-censored	Binary covariate	1.5	–0.032	0.203	0.042	0.220
		Uniform covariate	0.2	–0.010	0.089	0.008	0.089

Table S3: Bias, Stdev, MSE, and mean of standard error (SE) of the estimated coefficients for $n = 5000$ and three scenarios in case (a)when about 10% of individuals have experienced the event by the time of survey.

Scenario	Method	Covariates	True value	Bias	Stdev	MSE	Mean SE
Equal Reporting/Non-reporting	Observation-deletion	Binary covariate	1.5	0.146	0.125	0.037	0.131
		Uniform covariate	0.2	0.018	0.060	0.004	0.060
	Event-deletion	Binary covariate	1.5	–0.481	0.130	0.248	0.124
		Uniform covariate	0.2	–0.057	0.059	0.007	0.059
	Simple replacement	Binary covariate	1.5	–0.316	0.102	0.110	0.100
		Uniform covariate	0.2	–0.039	0.050	0.005	0.047
	Interval-censored	Binary covariate	1.5	–0.056	0.101	0.013	0.102
		Uniform covariate	0.2	–0.005	0.050	0.002	0.48
Reporting Dominance	Observation-deletion	Binary covariate	1.5	0.036	0.086	0.009	0.087
		Uniform covariate	0.2	0.006	0.038	0.001	0.037
	Event-deletion	Binary covariate	1.5	–0.373	0.108	0.151	0.082
		Uniform covariate	0.2	–0.044	0.040	0.004	0.037
	Simple replacement	Binary covariate	1.5	–0.334	0.103	0.122	0.078
		Uniform covariate	0.2	–0.036	0.039	0.003	0.035
	Interval-censored	Binary covariate	1.5	–0.033	0.086	0.008	0.086
		Uniform covariate	0.2	–0.003	0.037	0.001	0.036
Non-Reporting Dominance	Observation-deletion	Binary covariate	1.5	0.271	0.136	0.092	0.132
		Uniform covariate	0.2	0.040	0.056	0.005	0.055
	Event-deletion	Binary covariate	1.5	–0.967	0.130	0.952	0.122
		Uniform covariate	0.2	–0.107	0.055	0.014	0.055
	Simple replacement	Binary covariate	1.5	–0.831	0.096	0.699	0.077
		Uniform covariate	0.2	–0.083	0.044	0.009	0.036
	Interval-censored	Binary covariate	1.5	–0.080	0.097	0.016	0.096
		Uniform covariate	0.2	–0.006	0.038	0.002	0.038

Table S4: Bias, Stdev, MSE, and mean of standard error (SE) of the estimated coefficients for $n = 300$ and three scenarios in case (b) when about 30% of individuals have experienced the event by the time of survey.

Scenario	Method	Covariates	True value	Bias	Stdev	MSE	Mean SE
Equal Reporting/Non-reporting	Observation-deletion	Binary covariate	1.5	0.324	0.518	0.372	0.511
		Uniform covariate	0.2	0.020	0.202	0.041	0.190
	Event-deletion	Binary covariate	1.5	–1.443	0.522	2.354	0.434
		Uniform covariate	0.2	–0.178	0.175	0.062	0.172
	Simple replacement	Binary covariate	1.5	–1.276	0.429	1.812	0.263
		Uniform covariate	0.2	–0.164	0.171	0.05	0.114
	Interval-censored	Binary covariate	1.5	0.172	0.470	0.251	0.604
		Uniform covariate	0.2	0.007	0.172	0.030	0.180
Reporting Dominance	Observation-deletion	Binary covariate	1.5	0.102	0.309	0.105	0.309
		Uniform covariate	0.2	0.002	0.130	0.017	0.122
	Event-deletion	Binary covariate	1.5	–0.756	0.427	0.754	0.280
		Uniform covariate	0.2	–0.111	0.127	0.028	0.120
	Simple replacement	Binary covariate	1.5	–0.787	0.417	0.793	0.261
		Uniform covariate	0.2	–0.112	0.119	0.027	0.112
	Interval-censored	Binary covariate	1.5	0.072	0.312	0.102	0.331
		Uniform covariate	0.2	–0.002	0.129	0.017	0.129
Non-Reporting Dominance	Observation-deletion	Binary covariate	1.5	0.415	0.605	0.537	0.609
		Uniform covariate	0.2	0.039	0.228	0.053	0.224
	Event-deletion	Binary covariate	1.5	–1.455	0.577	2.449	0.507
		Uniform covariate	0.2	–0.179	0.210	0.076	0.199
	Simple replacement	Binary covariate	1.5	–1.256	0.428	1.761	0.253
		Uniform covariate	0.2	–0.160	0.170	0.054	0.110
	Interval-censored	Binary covariate	1.5	0.123	0.477	0.242	0.600
		Uniform covariate	0.2	0.006	0.167	0.028	0.180

Table S5: Bias, Stdev, MSE, and mean of standard error (SE) of the estimated coefficients for $n = 1000$ and three scenarios in case (b) when about 30% of individuals have experienced the event by the time of survey.

Scenario	Method	Covariates	True value	Bias	Stdev	MSE	Mean SE
Equal Reporting/Non-reporting	Observation-deletion	Binary covariate	1.5	0.258	0.265	0.137	0.260
		Uniform covariate	0.2	0.033	0.096	0.010	0.096
	Event-deletion	Binary covariate	1.5	–1.431	0.261	2.117	0.229
		Uniform covariate	0.2	–0.176	0.090	0.039	.092
	Simple replacement	Binary covariate	1.5	–1.290	0.216	1.711	0.140
		Uniform covariate	0.2	–0.162	0.086	0.034	0.061
	Interval-censored	Binary covariate	1.5	0.070	0.241	0.063	0.236
		Uniform covariate	0.2	0.007	0.082	0.007	0.083
Reporting Dominance	Observation-deletion	Binary covariate	1.5	0.068	0.167	0.033	0.164
		Uniform covariate	0.2	0.006	0.065	0.004	0.064
	Event-deletion	Binary covariate	1.5	–0.827	0.199	0.724	0.149
		Uniform covariate	0.2	–0.109	0.068	0.016	0.063
	Simple replacement	Binary covariate	1.5	–0.868	0.194	0.791	0.138
		Uniform covariate	0.2	–0.111	0.064	0.016	0.060
	Interval-censored	Binary covariate	1.5	0.023	0.172	0.030	0.167
		Uniform covariate	0.2	–0.001	0.064	0.004	0.064
Non-Reporting Dominance	Observation-deletion	Binary covariate	1.5	0.345	0.324	0.224	0.0305
		Uniform covariate	0.2	0.044	0.114	0.015	0.112
	Event-deletion	Binary covariate	1.5	–1.455	0.311	2.212	0.268
		Uniform covariate	0.2	–0.180	0.107	0.044	0.107
	Simple replacement	Binary covariate	1.5	–1.274	0.254	1.686	0.135
		Uniform covariate	0.2	–0.161	0.078	0.032	0.060
	Interval-censored	Binary covariate	1.5	0.038	0.253	0.065	0.243
		Uniform covariate	0.2	0.002	0.084	0.007	0.084

Table S6: Bias, Stdev, MSE, and mean of standard error (SE) of the estimated coefficients for $n = 5000$ and three scenarios in case (b) when about 30% of individuals have experienced the event by the time of survey.

Scenario	Method	Covariates	True value	Bias	Stdev	MSE	Mean SE
Equal Reporting/Non-reporting	Observation-deletion	Binary covariate	1.5	0.238	0.107	0.068	0.112
		Uniform covariate	0.2	0.032	0.043	0.003	0.042
	Event-deletion	Binary covariate	1.5	–1.419	0.105	2.023	0.101
		Uniform covariate	0.2	–0.178	0.040	0.033	0.041
	Simple replacement	Binary covariate	1.5	–1.296	0.092	1.688	0.062
		Uniform covariate	0.2	–0.164	0.037	0.028	0.027
	Interval-censored	Binary covariate	1.5	0.046	0.065	0.006	0.090
		Uniform covariate	0.2	0.005	0.026	0.001	0.030
Reporting Dominance	Observation-deletion	Binary covariate	1.5	0.049	0.067	0.007	0.072
		Uniform covariate	0.2	0.007	0.029	0.001	0.028
	Event-deletion	Binary covariate	1.5	–0.835	0.085	0.704	0.066
		Uniform covariate	0.2	–0.105	0.031	0.012	0.028
	Simple replacement	Binary covariate	1.5	–0.884	0.080	0.787	0.061
		Uniform covariate	0.2	–0.109	0.029	0.013	0.026
	Interval-censored	Binary covariate	1.5	0.002	0.068	0.005	0.073
		Uniform covariate	0.2	0.0003	0.028	0.001	0.028
Non-Reporting Dominance	Delete no recall	Binary covariate	1.5	0.311	0.127	0.113	0.130
		Uniform covariate	0.2	0.044	0.048	0.004	0.048
	Event-deletion	Binary covariate	1.5	–1.434	0.120	2.070	0.117
		Uniform covariate	0.2	–0.181	0.047	0.035	0.047
	Simple replacement	Binary covariate	1.5	–1.278	0.092	1.643	0.060
		Uniform covariate	0.2	–0.163	0.034	0.028	0.026
	Interval-censored	Binary covariate	1.5	0.0003	0.097	0.009	0.100
		Uniform covariate	0.2	0.001	0.035	0.001	0.035

S6 Two other examples

We utilized data on Myocardial infarction (MI) and osteoporosis/osteopenia from the CCSS as additional examples.

S6.1 Example S1: Myocardial infarction

In the CCSS dataset, the proportion of survivors who self-reported experiencing an MI is 1.2%. Among those reporting an MI, 14.5% did not report the age at onset, categorized as missing data. As indicated by Mulrooney et al. (2020), heart radiation dose, hypertension, and age at diagnosis were identified previously as risk factors associated with the MI . In our analysis, we used these predictors, specifying heart radiation dose as a categorical covariate (none, low, medium, and high dose), hypertension as a binary covariate, and age at primary cancer diagnosis as a continuous covariate. The results from the four methods in Section 3 of the main manuscript are summarizes in Table S7.

Table S7: Estimated hazard ratios and their 95% confidence intervals (CIs) of risk factors associated with the onset of MI

	Observation-deletion (N=25613)		Event-deletion (n=25656)		Simple replacement (n=25656)		Interval censored (n=25656)	
Parameter	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
No heart radiation	Ref.		Ref.		Ref.		Ref.	
Heart RT (0,15) Gy	1.12	(0.82, 1.53)	1.12	(0.82, 1.53)	1.11	(0.83, 1.48)	1.13	(0.83, 1.54)
Heart RT [15,35) Gy	2.74	(2.00, 3.75)	2.72	(1.99, 3.73)	2.65	(1.98, 3.54)	2.69	(1.97, 3.67)
Heart RT dose 35^{+} Gy	3.51	(1.84, 6.67)	3.52	(1.85, 6.68)	3.08	(1.63, 5.81)	3.16	(1.17, 8.50)
No hypertension	Ref.		Ref.		Ref.		Ref.	
Hypertension	4.21	(3.27, 5.43)	4.19	(3.25, 5.40)	3.84	(3.04, 4.85)	3.99	(3.12, 5.10)
Age at diagnosis	1.65	(1.45, 1.89)	1.66	(1.45, 1.89)	1.58	(1.40, 1.78)	1.57	(1.42, 1.73)

Our data reveals a broad range of post-treatment intervals for MI onset (among those reporting their onset age, the average onset age was 16.9, with the standard deviation of 9.2.) The incidence of self-reported MI within the CCSS cohort is relatively low (1.2%), with only 14.5% not report their onset age. Consequently, the performance of the four methods is quite similar, with the deletion-based approaches suggesting a slightly higher estimated risk associated with a heart radiation dose of 35 Gy or more compared to no heart radiation, in contrast to the other two methods.

S6.2 Example S2: Osteoporosis/osteopenia

Survivors of childhood cancer often confront challenges related to diminished bone mineral density and disrupted bone metabolism, rendering them more susceptible to heightened risks of fractures and skeletal complications. Within the CCSS data, 7.9% of survivors self-reported experiencing osteoporosis/osteopenia. Among them, 36.8% didn't report the age at onset, constituting missing data. For survivors who provided their onset age, the average onset age was 13.1, with a standard deviation of 9.5. Previous studies by Gawade et al. (2012); Gurney et al. (2014) identified risk factors such as brain radiation, any chemotherapy treatment, gender, and age at primary cancer diagnosis as the risk factors associated with osteoporosis/osteopenia. We used the same predictor in our analysis with brain radiation dose as a categorical covariate (none, low, medium, and high dose),

exposure to any chemotherapy treatment and gender as binary covariates, and age at diagnosis of primary cancer as a continuous covariate. Table S8 summarizes the findings.

Table S8: Estimated hazard ratios and their 95% confidence intervals (CIs) of risk factors associated with the onset of Osteoporosis/osteopenia

	Observation-deletion (N=24912)		Event-deletion (n=25656)		Simple replacement (n=25656)		Interval censored (n=25656)	
Parameter	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
No chemotherapy	Ref.		Ref.		Ref.		Ref.	
Chemotherapy	0.86	(0.77, 0.96)	0.87	(0.78, 0.97)	0.76	(0.69, 0.83)	0.78	(0.72, 0.86)
No brain radiation	Ref.		Ref.		Ref.		Ref.	
Brain RT (0,30) Gy	1.31	(1.13, 1.51)	1.27	(1.09, 1.47)	1.46	(1.31, 1.64)	1.57	(1.38, 1.78)
Brain RT [30,50) Gy	2.03	(1.52, 2.71)	1.90	(1.42, 2.54)	2.27	(1.83, 2.82)	2.37	(1.75, 3.22)
Brain RT 50^{+} Gy	2.40	(2.02, 2.85)	2.29	(1.93, 2.72)	2.84	(2.48, 3.24)	2.66	(2.26, 3.14)
Female	Ref.		Ref.		Ref.		Ref.	
Male	0.42	(0.37, 0.47)	0.43	(0.38, 0.48)	0.53	(0.48, 0.58)	0.52	(0.49, 0.55)
Age at diagnosis	1.24	(1.17, 1.31)	1.22	(1.16, 1.29)	1.24	(1.19, 1.30)	1.25	(1.21, 1.29)

With a high rate of missing onset age, around 40%, variations emerge in estimating hazard ratios for covariates across the different methods. For both the chemotherapy and brain radiation dose covariates, the hazard ratio estimates are attenuated towards null in the two deletion-based methods than the 'Simple replacement' and 'Interval-censored' methods, where the latter two methods give similar estimates. In our dataset, there is notable variability in the duration between five years post-diagnosis and up to 40 years thereafter for the manifestation of the condition among individuals who reported the onset age of osteoporosis/osteopenia. This variability aligns with the survey time range among those with the condition who did not report onset age, leading to comparable results in the 'Simple replacement' and 'Interval-censored' methods.

References

Anderson-Bergman, C. (2017). icenreg: Regression models for interval censored data in r. *Journal of Statistical Software*, 81(12):1–23.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics*, 59(3):570–579.

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

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42(4):845–854.

Gawade, P., Ness, K., Sharma, S., Li, Z., Srivastava, D., Spunt, S., Nottage, K., Krasin, M., Hudson, M., and Kaste, S. (2012). Association of bone mineral density with incidental renal stone in long-term survivors of childhood acute lymphoblastic leukemia. *Journal of Cancer Survivors.*, 6(4):388–397.

Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A., and Zaslavsky, A. M. (1998). A markov chain monte carlo em algorithm for analyzing interval-censored data under the cox proportional hazards model. *Biometrics*, pages 1498–1507.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media.

Gurney, J., Kadan-Lottick, N., Packer, R., Neglia, J., Sklar, C., Punyko, J., Stovall, M., Yasui, Y., Nicholson, H., Wolden, S., McNeil, D., Mertens, A., and Robison, L. (2003). Childhood cancer survivor study. endocrine and cardiovascular late effects among adult survivors of childhood brain tumors: Childhood cancer survivor study. *Cancer.*, 97(3):663–73.

Gurney, J., Kaste, S., Liu, W., Srivastava, D., Chemaitilly, W., Ness, K., Lanctot, J., Ojha, R., Nottage, KA. Wilson, C., Li, Z., Robison, L., and Hudson, M. (2014). Bone mineral density among long-term survivors of childhood acute lymphoblastic leukemia: Results from the st. jude lifetime cohort study. *Pediatric Blood Cancer.*, 61(7):1270–1276.

Leisenring, W., Mertens, A., Armstrong, G., Stovall, M., Neglia, J., Lanctot, J., Boice Jr, J., Whitton, J., and Yasui, Y. (2009). Pediatric cancer survivorship research: experience of the Childhood Cancer Survivor Study. *Journal of Clinical Oncology*, 27,:2319–27.

Li, J. and Ma, S. (2013). *Survival analysis in medicine and genetics*. CRC Press.

Mulrooney, D. A., Hyun, G., Ness, K. K., Ehrhardt, M. J., Yasui, Y., Duprez, D., Howell, R. M., Leisenring, W. M., Constine, L. S., Tonorezos, E., Gibson, T. M., Robison, L. L., Oeffinger, K. C., Hudson, M. M., and Armstrong, G. T. (2020). Major cardiac events for adult survivors of childhood cancer diagnosed between 1970 and 1999: report from the Childhood Cancer Survivor Study cohort. *British Medical Journal*, 368:2181–2189.

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science*, 3(4):425 – 441.

Robison, L., Armstrong, G., Boice, J., Chow, E., Davies, S., Donaldson, S., Green, D., Hammond, S., Meadows, A., Mertens, A., Mulvihill, J., Nathan, P., Neglia, J., Packer, R., Rajaraman, P., Sklar, C., Stovall, M., Strong, L., Yasui, Y., and Zeltzer, L. (2009). The childhood cancer survivor study: a National Cancer Institute-supported resource for outcome and intervention research. *Journal of Clinical Oncology*, 27,:2308–2318.

Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, 83(2):355–370.

Shao, F., Li, J., Ma, S., and Lee, M.-L. T. (2014). Semiparametric varying-coefficient model for interval censored data with a cured proportion. *Statistics in medicine*, 33(10):1700–1712.

Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data*. Springer, New York.

Taylor, J., Murray, S., and Hsu, C.-H. (2002). Survival estimation and testing via multiple imputation. *Statistics and Probability Letters.*, 58(3):221—232.

Wang, L., McMahan, C., Hudgens, M., and Qureshi, Z. (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics*, 1,(72):222–31.

Zhao, Y., Herring, A., Zhou, H., Ali, M., and Koch, G. (2014). A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring. *J Biopharm Stat.*, 24(2):229–53.