Lifetime Data Analysis
https://doi.org/10.1007/s10985-024-09621-2

# Cox model inference for relative hazard and pure risk from stratified weight-calibrated case-cohort data

Lola Etievant[1] · Mitchell H. Gail[1]

Received: 20 June 2023 / Accepted: 30 January 2024

## Abstract
The case-cohort design obtains complete covariate data only on cases and on a random sample (the subcohort) of the entire cohort. Subsequent publications described the use of stratification and weight calibration to increase efficiency of estimates of Cox model log-relative hazards, and there has been some work estimating pure risk. Yet there are few examples of these options in the medical literature, and we could not find programs currently online to analyze these various options. We therefore present a unified approach and R software to facilitate such analyses. We used influence functions adapted to the various design and analysis options together with variance calculations that take the two-phase sampling into account. This work clarifies when the widely used "robust" variance estimate of Barlow (Biometrics 50:1064–1072, 1994) is appropriate. The corresponding R software, CaseCohortCoxSurvival, facilitates analysis with and without stratification and/or weight calibration, for subcohort sampling with or without replacement. We also allow for phase-two data to be missing at random for stratified designs. We provide inference not only for log-relative hazards in the Cox model, but also for cumulative baseline hazards and covariate-specific pure risks. We hope these calculations and software will promote wider use of more efficient and principled design and analysis options for case-cohort studies.

**Keywords** Case-cohort · Weight calibration · Stratification · Variance estimation · Two-phase sampling · Influence functions

✉ Lola Etievant
lola.etievant@nih.gov

✉ Mitchell H. Gail
gailm@mail.nih.gov

[1] Division of Cancer Epidemiology and Genetics, Biostatistics Branch, National Cancer Institute, 9609 Medical Center Drive, Rockville, MD 20850-9780, USA

Published online: 02 April 2024

## 1 Introduction

Prentice (1986) described the case-cohort design for time-to-response outcomes, in which one obtains covariate information on all cases (those with the event) and on a random subcohort (which may include some cases) from the entire study cohort. Two great advantages of this design are that hard-to-measure covariates need only be obtained for the cases and subcohort, which is much smaller than the entire study cohort, and the data from the subcohort can be used for several different types of time-to-response outcomes. There have been subsequent refinements and extensions of this design. Barlow (1994) proposed a widely used "robust" variance estimator for log-relative hazards (RH) based on the sum of squared influences. Borgan et al. (2000) showed that a stratified case-cohort design had increased efficiency, and Samuelsen et al. (2007) and Gray (2009) noted that the "robust" variance estimate overestimated variances of log-relative hazard estimates with stratification when sampling without replacement. Breslow et al. (2009a, 2009b) proposed survey weight calibration to improve efficiency of case-cohort estimates of relative hazard. Although much of this literature focused on estimation of log-relative hazards, some authors considered estimation of cumulative baseline hazard and covariate-specific "pure" risk of an event (Chapters 16 and 17 in Borgan et al. 2017; Breslow and Lumley 2013; Gray 2009).

Sharp et al. (2014) noted variability in the analysis and reporting of 32 case-cohort studies from 24 major medical and epidemiological journals. None of these analyses used weight calibration, some used an inappropriate "robust" variance estimate with stratified data, and various methods were used for missing covariate information. Our informal review of subsequent case-cohort publications also indicates that stratification, weight calibration, a principled approach to missing subcohort data, and analysis of pure risk are underutilized. This may be partly due to difficulty understanding the highly technical and varied methodologic literature and to lack of convenient software.

To facilitate wider use of improved design and analysis options for case-cohort data, we unify the various analytic options above by presenting empirical influence functions for log-relative hazards and pure risk under a Cox proportional hazards model. These influence functions are adapted to the various design and analytic options above, and variance calculations acknowledge the phase-one sampling of the cohort from a superpopulation and the phase-two sampling of the subcohort. We develop software so that users can conveniently analyze case-cohort data with or without stratification and with or without weight calibration and can handle stratified case-cohort data with missing phase-two data.

We introduce notation in Sect. 2 and inference for the stratified case-cohort design in Sect. 3, which includes the unstratified design as a special case with one stratum. We describe weight calibration in Sect. 4, and methods for missing phase-two data in Sect. 5. We discuss current software in Sect. 6. Sections 7 and 8 present simulations and a data illustration, where we investigate the comparative efficiencies associated with stratification and weight calibration for hazards and covariate-specific pure risk, and how the "robust" variance estimate performs, with or without calibration. Concluding remarks are in Sect. 9. Most technical derivations and details are in Web Appendices.

## 2 Notation

We let $J$ be the number of strata in the whole cohort, $n^{(j)}$ be the number of subjects in stratum $j$, $j \in \{1, \dots, J\}$. Then $n = \sum_{j=1}^{J} n^{(j)}$ is the number of subjects in the whole cohort. We allow for right censoring and left truncation. We let $T_{i,j}$ be the event time (or age if the analysis is on the age scale) for subject $i$ in stratum $j$, and $C_{i,j}$ be the censoring time for subject $i$ in stratum $j$, $i \in \left\{1, \dots, n^{(j)}\right\}$, $j \in \{1, \dots, J\}$. Using the time-on-study scale, the at-risk indicator for subject $i$ in stratum $j$ is $Y_{i,j}(t) = I\left(\tilde{T}_{i,j} \geq t\right)$, with $\tilde{T}_{i,j} = \min(T_{ij}, C_{ij})$. Using the age scale, $Y_{i,j}(t) = I\left(\tilde{T}_{i,j} \geq t > E_{i,j}\right)$, with $E_{i,j}$ the entry age for subject $i$ in stratum $j$. Let $\tau$ the maximum follow-up time or maximum age for analyses on the age scale. With $N_{i,j}(t) = I\left(T_{i,j} \leq t, T_{ij} \leq C_{ij}\right)$ indicating an observed event before or at time/age $t$ after study entry, $\mathrm{d}N_{i,j}(t)$ indicates if individual $i$ in stratum $j$ fails (has the event) at time/age $t$. Finally, we let $\boldsymbol{X}_{i,j}$ be a vector of $p$ baseline covariates for subject $i$ in stratum $j$; $\boldsymbol{X}_{i,j}$ includes stratum indicators or stratum determinants.

We assume that failure follows the Cox proportional hazards model with hazard function $\lambda(t) = \lambda_0(t) \exp\left(\boldsymbol{\beta}'\boldsymbol{X}\right)$, for covariates $\boldsymbol{X}$, and where $\lambda_0(t)$ is a baseline hazard function, i.e., the hazard for an individual with $\boldsymbol{X} = \mathbf{0}$. We further assume that $\lambda_0(t)$ is homogeneous across strata and we let $\Lambda_0(t) = \int\limits_0^t \lambda_0(s)\mathrm{d}s$ denote the cumulative baseline hazard.

Estimation from complete cohort data is reviewed in Web Appendix A.1 with corresponding influence functions in Web Appendix A.2.

## 3 Stratified case-cohort

### 3.1 Estimation of relative hazard, cumulative baseline hazard and pure risk

We assume that a fixed number of individuals, $m^{(j)}$, is sampled from stratum $j$ (of size $n^{(j)}$) in the cohort, without replacement and independently of case status, $j \in \{1, \dots, J\}$. Sampling is performed independently across strata. The subcohort includes all the sampled subjects from the $J$ strata. In addition, we sample all the cases in the cohort, some of whom may have been included in the subcohort. All of these individuals constitute the stratified case-cohort, that we also call the *phase-two sample*, because it is a subset of the cohort, which is regarded as a *phase-one sample* from a super-population. We let

$\xi_{i,j}$ be the sampling indicator of individual $i$ in stratum $j$ and $w_{i,j} = \begin{cases} \frac{n^{(j)}}{m^{(j)}} & \text{if } i \text{ is a non-case in stratum } j \\ 1 & \text{if } i \text{ is a case in stratum } j \end{cases}$ be his/her known design weight, $i \in \left\{1, \dots, n^{(j)}\right\}$, $j \in \{1, \dots, J\}$. We assume that some of the covariates in $\boldsymbol{X}$ are only measured in the phase-two sample; we call these "phase-two covariates". The stratum indicators are known for all members of the cohort and are not phase-two covariates. Because we sample all cases, $\xi_{i,j} w_{i,j} = 1$ for cases. Non-stratified case-cohort data correspond to the special case $J = 1$.

An estimate of the log-relative hazard $\boldsymbol{\beta}$ is obtained by solving the estimating equation

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \int_{t} \left\{ \boldsymbol{X}_{i,j} - \frac{\boldsymbol{S}_1(t;\boldsymbol{\beta})}{S_0(t;\boldsymbol{\beta})} \right\} \mathrm{d}N_{i,j}(t) = 0, \tag{1}$$

with

$$S_0(t;\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{k=1}^{n^{(j)}} w_{k,j} \xi_{k,j} Y_{k,j}(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{X}_{k,j}\right), \tag{2}$$

$$\boldsymbol{S}_1(t;\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{k=1}^{n^{(j)}} w_{k,j} \xi_{k,j} Y_{k,j}(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{X}_{k,j}\right) \boldsymbol{X}_{k,j}, \tag{3}$$

and we also define

$$\boldsymbol{S}_2(t;\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{k=1}^{n^{(j)}} w_{k,j} \xi_{k,j} Y_{k,j}(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{X}_{k,j}\right) \boldsymbol{X}_{k,j} \boldsymbol{X}'_{k,j}. \tag{4}$$

Let $\widehat{\boldsymbol{\beta}}$ denote this solution. We could write Eq. (1) as $\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \int_t \left\{ \boldsymbol{X}_{i,j} - \frac{\boldsymbol{S}_1(t;\boldsymbol{\beta})}{S_0(t;\boldsymbol{\beta})} \right\}$ $\xi_{i,j} w_{i,j} \mathrm{d}N_{i,j}(t) = 0$, because $\xi_{i,j} w_{i,j} = 1$ for cases; this form would be useful if cases were subsampled (see Sect. 9). We then estimate the baseline hazard point mass at time $t$ non-parametrically (Breslow 1974) by

$$\mathrm{d}\widehat{\Lambda}_0\left(t;\widehat{\boldsymbol{\beta}}\right) \equiv \mathrm{d}\widehat{\Lambda}_0(t) = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \mathrm{d}N_{i,j}(t)}{S_0\left(t, \widehat{\boldsymbol{\beta}}\right)}, \tag{5}$$

the cumulative baseline hazard up to time $t$ by

$$\widehat{\Lambda}_0\left(t;\widehat{\boldsymbol{\beta}}, \widehat{\lambda}_0\right) \equiv \widehat{\Lambda}_0(t) = \int_0^t \mathrm{d}\widehat{\Lambda}_0(s), \tag{6}$$

and the pure covariate-specific risk for profile $\boldsymbol{x}$ in the interval $(\tau_1, \tau_2]$ by

$$\hat{\pi}\left(\tau_1, \tau_2; \boldsymbol{x}, \hat{\boldsymbol{\beta}}, \mathrm{d}\hat{\Lambda}_0\right) \equiv \hat{\pi}\left(\tau_1, \tau_2; \boldsymbol{x}\right) = 1 - \exp\left\{-\int_{\tau_1}^{\tau_2} \exp\left(\hat{\boldsymbol{\beta}}' \boldsymbol{x}\right) \mathrm{d}\hat{\Lambda}_0(s)\right\}. \tag{7}$$

In Sects. 3.3 and 4.3 we show how to use influence functions to estimate the variance of $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_0(t)$ for a fixed $t$.

### 3.2 Influence functions

As described in Deville (1999), survey samplers often compute the variance of a statistic $\hat{\boldsymbol{\theta}}$ with expectation $\boldsymbol{\theta}$ by using the linear approximation $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \sum \boldsymbol{\Delta}_i + \boldsymbol{R}$, where the remainder $\boldsymbol{R}$ is of smaller order than $\hat{\boldsymbol{\theta}}$ and the summation is over the sample units indexed by $i$. The variance of $\hat{\boldsymbol{\theta}}$ can therefore be calculated as $\mathrm{var}\left(\sum \boldsymbol{\Delta}_i\right)$. The $\boldsymbol{\Delta}_i$ are called Taylor deviates or influences (in other literature, e.g. Tsiatis (2006), the $\boldsymbol{\Delta}_i$ divided by the sample size are called influences). Using the calculus for Taylor deviates described in Deville (1999) and Graubard and Fears (2005), we calculated the influences in this Section. The $\boldsymbol{\Delta}_i$ are theoretical quantities that depend on unknown parameters, but substituting consistent estimates of these parameters to produce "empirical" influences still yields asymptotically consistent variance estimates (Deville 1999).

We let $\boldsymbol{\Delta}_{i,j}\left(\hat{\boldsymbol{\theta}}\right)$ denote the empirical influence of subject $i$ in stratum $j$ on $\hat{\boldsymbol{\theta}}$ from the set $\left\{\hat{\boldsymbol{\beta}}, \mathrm{d}\hat{\Lambda}_0(t), \hat{\Lambda}_0(t), \hat{\pi}\left(\tau_1, \tau_2; \boldsymbol{x}\right)\right\}$, $i \in \left\{1, \dots, n^{(j)}\right\}, j \in \{1, \dots, J\}$. From these influences, we can estimate the covariance matrix of $\hat{\boldsymbol{\theta}}$ as (Graubard and Fears 2005)

$$\mathrm{var}\left(\hat{\boldsymbol{\theta}}\right) \approx \mathrm{var}\left\{\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \boldsymbol{\Delta}_{i,j}\left(\hat{\boldsymbol{\theta}}\right)\right\}. \tag{8}$$

Following Graubard and Fears (2005) and Section 4.6 in Pfeiffer and Gail (2017), we show in Web Appendix B.1 that $\boldsymbol{\Delta}_{i,j}\left(\hat{\boldsymbol{\theta}}\right) = \xi_{i,j} w_{i,j} \boldsymbol{IF}^{(2)}_{i,j}\left(\hat{\boldsymbol{\theta}}\right)$, where

$$\begin{aligned} \boldsymbol{IF}^{(2)}_{i,j}\left(\hat{\boldsymbol{\beta}}\right) = &\left[\sum_{l=1}^{J} \sum_{k=1}^{n^{(l)}} \int_t \left\{\frac{\boldsymbol{S}_2\left(t; \hat{\boldsymbol{\beta}}\right)}{S_0\left(t; \hat{\boldsymbol{\beta}}\right)} - \frac{\boldsymbol{S}_1\left(t; \hat{\boldsymbol{\beta}}\right) \boldsymbol{S}_1\left(t; \hat{\boldsymbol{\beta}}\right)'}{S_0\left(t; \hat{\boldsymbol{\beta}}\right)^2}\right\} \mathrm{d}N_{k,l}(t)\right]^{-1} \left[\int_t \left\{\boldsymbol{X}_{i,j} - \frac{\boldsymbol{S}_1\left(t; \hat{\boldsymbol{\beta}}\right)}{S_0\left(t; \hat{\boldsymbol{\beta}}\right)}\right\}\right. \\ &\left.\left\{\mathrm{d}N_{i,j}(t) - \frac{Y_{i,j}(t) \exp\left(\hat{\boldsymbol{\beta}}' \boldsymbol{X}_{i,j}\right) \sum_{l=1}^{J} \sum_{k=1}^{n^{(l)}} \mathrm{d}N_{k,l}(t)}{S_0\left(t; \hat{\boldsymbol{\beta}}\right)}\right\}\right], \end{aligned} \tag{9}$$

$$\begin{aligned} \boldsymbol{IF}_{i,j}^{(2)}\left\{\mathrm{d}\widehat{\Lambda}_0(t)\right\} &= \left\{S_0\left(t;\widehat{\boldsymbol{\beta}}\right)\right\}^{-1}\left\{\mathrm{d}N_{i,j}(t) - \mathrm{d}\widehat{\Lambda}_0(t)Y_{i,j}(t)\exp\left(\widehat{\boldsymbol{\beta}}'\boldsymbol{X}_{i,j}\right)\right\} \\ &\quad - \left\{S_0\left(t;\widehat{\boldsymbol{\beta}}\right)\right\}^{-1}\mathrm{d}\widehat{\Lambda}_0(t)\boldsymbol{S}_1\left(t;\widehat{\boldsymbol{\beta}}\right)'\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\beta}}\right), \end{aligned} \tag{10}$$

$$IF_{i,j}^{(2)}\left\{\int_{\tau_1}^{\tau_2}\mathrm{d}\widehat{\Lambda}_0(t)\right\} = \int_{\tau_1}^{\tau_2} IF_{i,j}^{(2)}\left\{\mathrm{d}\widehat{\Lambda}_0(t)\right\}, \tag{11}$$

and

$$\begin{aligned} IF_{i,j}^{(2)}\left\{\widehat{\pi}\left(\tau_1,\tau_2;\boldsymbol{x}\right)\right\} &= \left\{\frac{\partial\widehat{\pi}\left(\tau_1,\tau_2;\boldsymbol{x}\right)}{\partial\boldsymbol{\beta}}_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right\}\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\beta}}\right) \\ &\quad + \left[\frac{\partial\widehat{\pi}\left(\tau_1,\tau_2;\boldsymbol{x}\right)}{\partial\left\{\int_{\tau_1}^{\tau_2}\mathrm{d}\Lambda_0(t)\right\}}_{|\mathrm{d}\Lambda_0(t)=\mathrm{d}\widehat{\Lambda}_0(t)}\right] IF_{i,j}^{(2)}\left\{\int_{\tau_1}^{\tau_2}\mathrm{d}\widehat{\Lambda}_0(t)\right\}. \end{aligned} \tag{12}$$

Equations (9)–(12) depend on "phase-two covariates". Hence, we use the superscript 2 in $\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)$.

### 3.3 Variance decomposition and estimation from influence functions

The variance $\mathrm{var}\left(\widehat{\boldsymbol{\theta}}\right) \approx \mathrm{var}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\boldsymbol{\Delta}_{i,j}\left(\widehat{\boldsymbol{\theta}}\right)\right\}$ can be decomposed as

$$\mathrm{var}\left[\mathrm{E}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\boldsymbol{\Delta}_{i,j}\left(\widehat{\boldsymbol{\theta}}\right)|\boldsymbol{C}_1\right\}\right] + \mathrm{E}\left[\mathrm{var}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\boldsymbol{\Delta}_{i,j}\left(\widehat{\boldsymbol{\theta}}\right)|\boldsymbol{C}_1\right\}\right], \tag{13}$$

where $\boldsymbol{C}_1$ denotes the information from the whole cohort. The first component accounts for sampling the cohort from the "superpopulation" (phase-one component of variance), whereas the second component accounts for sampling the subcohort from the cohort (phase-two component of variance).

We let $w_{i,k,j}$ and $\sigma_{i,k,j}$ denote $\mathrm{E}\left(\xi_{i,j}\xi_{k,j}|\boldsymbol{C}_1\right)^{-1}$ and $\mathrm{cov}\left(\xi_{i,j},\xi_{k,j}|\boldsymbol{C}_1\right)$, respectively, $i,k \in \left\{1,\ldots,n^{(j)}\right\}$, $j \in \{1,\ldots,J\}$; they are specified below. We know $w_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)$ is fixed conditional on $\boldsymbol{C}_1$ and $\mathrm{E}\left(\xi_{i,j}w_{i,j}|\boldsymbol{C}_1\right) = 1$. Thus $\mathrm{var}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\boldsymbol{\Delta}_{i,j}\left(\widehat{\boldsymbol{\theta}}\right)\right\} =$ $\mathrm{var}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)\right\} + \mathrm{E}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\sum_{k=1}^{n^{(j)}}\sigma_{i,k,j}w_{i,j}w_{k,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{k,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)'\right\}$.

Because $\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)'$ and $\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{k,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)'$ can only be computed if individuals $i$ and $k$ in stratum $j$ are in the phase-two sample, we weight the contributi ons from the individuals in the phase-two sample by the “marginal” and “joint” design weights, $w_{i,j}$ and $w_{i,k,j}$, to estimate $\mathrm{var}\left(\hat{\boldsymbol{\theta}}\right)$ by

$$\frac{n}{n-1}\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\xi_{i,j}w_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)' + \sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\sum_{k=1}^{n^{(j)}}w_{i,k,j}\sigma_{i,k,j}w_{i,j}w_{k,j}\xi_{i,j}\xi_{k,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{k,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)'. \tag{14}$$

Following Barlow (1994), the “robust” variance estimate would be

$$\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\boldsymbol{\Delta}_{i,j}\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{\Delta}_{i,j}\left(\hat{\boldsymbol{\theta}}\right)' = \sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\xi_{i,j}w_{i,j}w_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)'. \tag{15}$$

With stratified data, Eq. (15) is often too large (see also Sect. 7 and Web Appendix D.2). Equation (15) minus Eq. (14) is

$$\frac{1}{n-1}\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\xi_{i,j}w_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)' + \sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\sum_{\substack{k=1\\k\neq i}}^{n^{(j)}}w_{i,k,j}\sigma_{i,k,j}w_{i,j}w_{k,j}\xi_{i,j}\xi_{k,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{k,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)'. \tag{16}$$

Because we sample *without replacement* in each stratum, we have $w_{i,k,j} = \frac{n^{(j)}\left(n^{(j)}-1\right)}{m^{(j)}\left(m^{(j)}-1\right)}$ if individuals $i$ and $k$ in stratum $j$ are both non-cases, and $w_{i,k,j} = w_{i,j} \times w_{k,j}$ otherwise, $i,k \in \left\{1,\dots,n^{(j)}\right\}$, $k \neq i$, $j \in \{1,\dots,J\}$. Recall that $w_{i,i,j} = w_{i,j} = \frac{n^{(j)}}{m^{(j)}}$ if individual $i$ in stratum $j$ is a non-case, and $w_{i,j} = 1$ if individual $i$ in stratum $j$ is a case. Then $\sigma_{i,k,j} = \frac{m^{(j)}}{n^{(j)}}\frac{m^{(j)}-1}{n^{(j)}-1} - \left(\frac{m^{(j)}}{n^{(j)}}\right)^2$ if individuals $i$ and $k$ in stratum $j$ are both non-cases, and $\sigma_{i,k,j} = 0$ otherwise, $i,k \in \left\{1,\dots,n^{(j)}\right\}$, $k \neq i$, $j \in \{1,\dots,J\}$. Similarly, if individual $i$ in stratum $j$ is a non-case, then $\sigma_{i,i,j} \equiv \sigma_{i,j} = \frac{m^{(j)}}{n^{(j)}}\left(1 - \frac{m^{(j)}}{n^{(j)}}\right)$, and $\sigma_{i,j} = 0$ otherwise. As a result, only the sampled non-cases contribute to the phase-two component of the variance in Eq. (14). For sampling *with replacement* (i.e., Bernoulli sampling), individuals are sampled independently of each other. Then $w_{i,k,j} = w_{i,j} \times w_{k,j}$, and $\sigma_{i,k,j} = 0$ for any pair $(i,k)$ of distinct individuals in stratum $j$, $i,k \in \left\{1,\dots,n^{(j)}\right\}$, $k \neq i$, $j \in \{1,\dots,J\}$. In that case, the difference between the “robust” variance estimate in Eqs. (15) and (14) reduces to $\frac{1}{n-1}\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\xi_{i,j}w_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)'$, which is negligible compared to Eq. (14) in large cohorts.

The variance estimate in Eq. (14) is asymptotically equivalent to that of Lin (2000) for $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_0(t)$ (Lin did not consider covariate-specific pure risks). The second component in Eq. (14) is precisely equal to the terms Lin (2000) used to estimate the phase-two component of the variance for $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_0(t)$. To estimate the phase-one component of variance of $\hat{\boldsymbol{\beta}}$, Lin (2000) used the inverse of the observed information matrix for $\boldsymbol{U}(\boldsymbol{\beta})$, whereas the weighted sum of $\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{IF}_{i,j}^{(2)}\left(\hat{\boldsymbol{\beta}}\right)'$ in the first component in Eq. (14) is a “sandwich estimate” of this quantity that is consistent for it. In addition, relying on the influences to estimate the phase-one component of the variance allows for an easy extension to other designs and analytic options (e.g., using calibrated weights). The estimated phase-one component of variance of $\widehat{\Lambda}_0(t)$ on page 43 of Lin (2000) consists of two parts, as in Andersen and Gill (1982), that correspond to the two terms in $IF^{(2)}_{i,j}\left\{\int_{\tau_1}^{\tau_2} \mathrm{d}\widehat{\Lambda}_0(t)\right\}$ obtained from Eqs. (10) and (11). The conditional expectation of the second term is zero given the phase-one data, proving that the two components are uncorrelated. The weighted sum of the first term squared equals that in Lin (2000). The weighted sum of the cross-products of the second term is a sandwich estimate of the second quantity estimated in Lin (2000), who instead used the observed information matrix for $\boldsymbol{U}(\boldsymbol{\beta})$ in the calulation. See also Web Appendix B.2 for comparison with the estimate of $\mathrm{var}\left(\widehat{\boldsymbol{\beta}}\right)$ by Samuelsen et al. (2007). We note that our influence function-based variance estimates performed well in simulations (see Sect. 7).

### 3.4 Asymptotic normality

We assume that normed estimates of $\widehat{\boldsymbol{\beta}}$, $\widehat{\Lambda}_0(t)$ and $\widehat{\pi}\left(\tau_1, \tau_2; \boldsymbol{x}\right)$ are normally distributed for fixed $t$, $\tau_1$ and $\tau_2$. This assumption is supported by nominal coverage of confidence intervals in our simulations. From finite sampling theory, Borgan et al. (2000) and Lin (2000) argued that certain phase-two normed sums were asymptotically normally distributed conditional on the phase-one data with covariances that did not depend on the phase-one data, implying that $n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ was asymptotically Normal unconditionally. Assuming an additional tightness condition, Lin (2000) proved that $n^{\frac{1}{2}}\left\{\widehat{\Lambda}_0(t) - \Lambda_0(t)\right\}$ converged to a Gaussian process. Hence $n^{1/2}\left\{\widehat{\pi}\left(\tau_1, \tau_2; \boldsymbol{x}\right) - \pi\left(\tau_1, \tau_2; \boldsymbol{x}\right)\right\}$ converges to normality for fixed $\tau_1$ and $\tau_2$. The tightness condition was not proved but thought to hold for stratified designs.

## 4 Calibration of the design weights

### 4.1 Calibration and choice of auxiliary variables

Breslow et al. (2009a, 2009b) advocated “weight calibration” to improve the efficiency of case-cohort studies. First, one identifies auxiliary variables that are highly correlated with the influences on $\widehat{\boldsymbol{\theta}}$ and are known for the entire cohort. Then one perturbs the design weights to obtain calibrated weights that are close to the design weights but for which the observed sums of auxiliary variables in the phase-one sample equals the weighted sums in the phase-two sample with the calibrated weights. To obtain auxiliary variables, we follow Shin et al.

(2020). First, we use weighted regression in the phase-two sample to estimate the expected value of phase-two covariates given phase-one data. These expectations are used to impute the phase-two covariates for all members of the cohort, including those with measured phase-two covariates. The phase-one data used for imputation may consist of covariates in $\boldsymbol{X}$ and of phase-one proxies of the phase-two covariates that are measured on all cohort members. The auxiliary variables are (i) the influences for the log-relative hazard parameters estimated from the Cox model with imputed cohort data; and (ii) the products of follow-up time on the interval for which pure risk is to be estimated times the estimated relative hazard for the imputed cohort data, where the log-relative hazard parameters are estimated from the Cox model with case-cohort data and weights calibrated with (i). To standardize the weights, we also calibrate against (iii) a variable that is identically equal to 1. Calibration of the design weights against (i) alone was proposed by Breslow et al. (2009a, 2009b) to improve efficiency of case-cohort estimates of log-relative hazard. Shin et al. (2020) extended the work of Breslow et al. (2009a, 2009b) and proposed calibrating against (i) + (ii) + (iii) to improve efficiency of log-relative hazard and pure risk estimates under the nested case–control design. Additional details are in Web Appendix C.1; see also Breslow et al. (2009a) and Shin et al. (2020). Other auxiliary variables have been proposed for $\widehat{\Lambda}_0(t)$ (Breslow and Lumley 2013), but in unreported simulations, the proposal by Shin et al. performed better; see also Web Appendix C.3.

We let $\boldsymbol{A}_{i,j}$ be the vector of $q$ auxiliary variables for individual $i$ in stratum $j$, with calibrated weights $w^*_{i,j} = w_{i,j} \exp\left(\widehat{\boldsymbol{\eta}}' \boldsymbol{A}_{i,j}\right)$, $i \in \left\{1, \dots, n^{(j)}\right\}$, $j \in \{1, \dots, J\}$, that are obtained by solving $\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \left\{\xi_{i,j} w_{i,j} \exp\left(\boldsymbol{\eta}' \boldsymbol{A}_{i,j}\right) \boldsymbol{A}_{i,j} - \boldsymbol{A}_{i,j}\right\} = 0$ for $\widehat{\boldsymbol{\eta}}$. See Web Appendix C.1.

## 4.2 Estimation of relative hazard, cumulative baseline hazard and pure risk using calibrated weights

An estimate of $\boldsymbol{\beta}$ solves $\boldsymbol{U}^*(\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \int_t \xi_{i,j} w^*_{i,j} \left\{ \boldsymbol{X}_{i,j} - \frac{\boldsymbol{S}^*_1(t;\widehat{\boldsymbol{\eta}},\boldsymbol{\beta})}{S^*_0(t;\widehat{\boldsymbol{\eta}},\boldsymbol{\beta})} \right\} \mathrm{d}N_{i,j}(t) = 0$, where $S^*_0\left(t;\widehat{\boldsymbol{\eta}}, \boldsymbol{\beta}\right)$, $\boldsymbol{S}^*_1\left(t;\widehat{\boldsymbol{\eta}}, \boldsymbol{\beta}\right)$ and $\boldsymbol{S}^*_2\left(t;\widehat{\boldsymbol{\eta}}, \boldsymbol{\beta}\right)$ are obtained from Eqs. (2)–(4) with $w^*_{k,j}$ replacing $w_{k,j}$. Letting $\widehat{\boldsymbol{\beta}}^*\left(\widehat{\boldsymbol{\eta}}\right) \equiv \widehat{\boldsymbol{\beta}}^*$, we estimate the baseline hazard point mass at time $t$, $\mathrm{d}\widehat{\Lambda}^*_0\left(t;\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\beta}}^*\right) \equiv \mathrm{d}\widehat{\Lambda}^*_0(t)$, the cumulative baseline hazard up to time $t$, $\widehat{\Lambda}^*_0\left(t;\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\beta}}^*\right) \equiv \widehat{\Lambda}^*_0(t)$, and the pure risk for profile $\boldsymbol{x}$ in the interval $(\tau_1, \tau_2]$, $\widehat{\pi}^*\left(\tau_1, \tau_2;\boldsymbol{x}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\beta}}^*, \mathrm{d}\widehat{\Lambda}^*_0\right) \equiv \widehat{\pi}^*\left(\tau_1, \tau_2;\boldsymbol{x}\right)$, from Eqs. (5)–(7) with $S^*_0\left(t;\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\beta}}^*\right)$ and $\widehat{\boldsymbol{\beta}}^*$ replacing $S_0\left(t, \widehat{\boldsymbol{\beta}}\right)$ and $\widehat{\boldsymbol{\beta}}$. We do not calibrate the case weights in the numerator of the Breslow estimator because the event times are known for all cohort members (Breslow and Wellner 2007; Pugh et al. 1993; Shin et al. 2020).

### 4.3 Variance estimation from influence functions

We let $\boldsymbol{\Delta}_{i,j}\left(\widehat{\boldsymbol{\theta}}^*\right)$ denote the influence of individual $i$ in stratum $j$ on one of the parameters $\widehat{\boldsymbol{\theta}}^*$ from the set $\left\{\hat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\beta}}^*, \mathrm{d}\widehat{\Lambda}_0^*(t), \widehat{\Lambda}_0^*(t), \widehat{\pi}^*\left(\tau_1, \tau_2; \boldsymbol{x}\right)\right\}$, $i \in \left\{1, \ldots, n^{(j)}\right\}$, $j \in \{1, \ldots, J\}$, and use $\mathrm{var}\left(\widehat{\boldsymbol{\theta}}^*\right) \approx \mathrm{var}\left\{\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \boldsymbol{\Delta}_{i,j}\left(\widehat{\boldsymbol{\theta}}^*\right)\right\}$. Following Shin et al. (2020) we can show that $\boldsymbol{\Delta}_{i,j}\left(\widehat{\boldsymbol{\theta}}^*\right) = \boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^*\right) + \xi_{i,j} w_{i,j} \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right)$. The superscript 1 indicates that $\boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^*\right)$ depends only on variables measured on all cohort members. If individual $i$ in stratum $j$ is not in the phase-two sample, $\xi_{i,j} w_{i,j} \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right)$ is zero, but such an individual has an influence on $\hat{\boldsymbol{\eta}}$ and hence on $\hat{\boldsymbol{\theta}}^*$ through $\boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^*\right)$. Explicit forms of $\boldsymbol{IF}_{i,j}^{(s)}\left(\widehat{\boldsymbol{\theta}}^*\right)$, $s \in \{1, 2\}$, are in Appendix 1 and derived in Web Appendix C.2.

Because $\boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}\right)$ is fixed conditional on $\boldsymbol{C}_1$, a decomposition similar to Eq. (13) yields

$$\begin{aligned} \mathrm{var}\left\{\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \boldsymbol{\Delta}_{i,j}\left(\widehat{\boldsymbol{\theta}}^*\right)\right\} = {} & \mathrm{var}\left\{\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^*\right) + \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right)\right\} \\ & + \mathrm{E}\left\{\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \sum_{k=1}^{n^{(j)}} \sigma_{i,k,j} w_{i,j} w_{k,j} \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right) \boldsymbol{IF}_{k,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right)'\right\}, \end{aligned} \tag{17}$$

which can be estimated by

$$\begin{aligned} & \frac{n}{n-1} \sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \left\{\boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^*\right) \boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^*\right)\prime + 2\xi_{i,j} w_{i,j} \boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^*\right) \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right)' \right. \\ & \left. + \xi_{i,j} w_{i,j} \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right) \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right)'\right\} \\ & + \sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \sum_{k=1}^{n^{(j)}} w_{i,k,j} \sigma_{i,k,j} \xi_{i,j} \xi_{k,j} w_{i,j} w_{k,j} \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right) \boldsymbol{IF}_{k,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right)'. \end{aligned} \tag{18}$$

Finally, the robust variance estimate (Barlow 1994) is

$$\begin{aligned} \sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \boldsymbol{\Delta}_{i,j}\left(\widehat{\boldsymbol{\theta}}^*\right) \boldsymbol{\Delta}_{i,j}\left(\widehat{\boldsymbol{\theta}}^*\right)' = {} & \sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \left\{\boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^*\right) \boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^*\right)' \right. \\ & \left. + 2\xi_{i,j} w_{i,j} \boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^*\right) \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right)' + \xi_{i,j} w_{i,j} w_{i,j} \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right) \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^*\right)'\right\}, \end{aligned} \tag{19}$$

and the difference between Eqs. (18) and (19) is

$$\frac{1}{n-1}\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\left\{\boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^{*}\right)\boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^{*}\right)' + 2\xi_{i,j}w_{i,j}\boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\theta}}^{*}\right)\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^{*}\right)' \right.$$
$$\left. +\xi_{i,j}w_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^{*}\right)\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^{*}\right)'\right\} \tag{20}$$
$$+\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\sum_{\substack{k=1\\k\neq i}}^{n^{(j)}} w_{i,k,j}\sigma_{i,k,j}w_{i,j}w_{k,j}\xi_{i,j}\xi_{k,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^{*}\right)\boldsymbol{IF}_{k,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^{*}\right)'.$$

For individuals $i$ and $k$ in stratum $j$ such that $\sigma_{i,k,j}$ and $\sigma_{i,j}$ are non-zero (i.e., non-cases), $\xi_{i,j}w_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^{*}\right)$ and $\xi_{k,j}w_{k,j}\boldsymbol{IF}_{k,j}^{(2)}\left(\widehat{\boldsymbol{\theta}}^{*}\right)$ are weighted residuals from a weighted linear regression on the auxiliary variables; see Web Appendix C.3. With good auxiliary variables for calibration, one can expect the phase-two component of the variance and hence the total variance to be smaller, and the difference in Eq. (20) to be smaller than the difference in Eq. (16); see also Chapter 17 in Borgan et al. (2017).

See Web Appendix I for a summary of the steps for parameter and variance estimation with calibrated weights.

## 5 Missing data

### 5.1 Notation

Covariate information may be missing for individuals in phase-two. For example, stored blood samples from individuals in phase-two could have been previously used or lost. We assume such covariates are missing at random and we regard the set of individuals with complete covariate data as a *phase-three sample*. More precisely, let $V_{i,j}$ be the phase-three sampling indicator for subject $i$ in stratum $j$, $i \in \left\{1,\dots,n^{(j)}\right\}$, $j \in \{1,\dots,J\}$; we assume the Bernoulli indicators $V_{i,j}$ are mutually independent and independent of the phase-two indicators, $\xi_{i,j}$. Let $w_{i,j}^{(3)} \equiv \frac{1}{\pi_{i,j}^{(3)}}$ be the phase-three design weight, where $\pi_{i,j}^{(3)}$ is the phase-three design sampling probability. The overall sampling design weight of subject $i$ in stratum $j$ is $w_{i,j} = w_{i,j}^{(2)} \times w_{i,j}^{(3)}$.

The phase-three sampling probabilities may differ in $J^{(3)}$ exclusive and exhaustive subsets (phase-three strata) of the population that need not coincide with the $J$ phase-two strata. For example, cases may have a different probability of missingness from non-cases. Nonetheless, we index the members of the cohort as in Sect. 3. Web Appendix F describes analysis when the phase-three sampling probabilities are known. However, the $\pi_{i,j}^{(3)}$ are usually unknown and need to be estimated (Sect. 5.2).

### 5.2 Weight estimation

When the $\pi_{i,j}^{(3)}$ are unknown, $w_{i,j}^{(3)}$ can be estimated as follows, $i \in \left\{1, \dots, n^{(j)}\right\}$, $j \in \{1, \dots, J\}$. The $V_{i,j}$ are known for all members of the phase-two sample, and let $\boldsymbol{B}_{i,j}$ be a $J^{(3)} \times 1$ vector of indicator variables that take value 1 if subject $i$ in (phase-two) stratum $j$ is in the corresponding phase-three stratum, and 0 otherwise. Let $\exp\left(\widetilde{\boldsymbol{\gamma}}\right)$ be the vector of $J^{(3)}$ estimated phase-three sampling weights that are obtained by solving the estimating equation $\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \xi_{i,j} \boldsymbol{B}_{i,j} - \exp\left(\boldsymbol{\gamma}' \boldsymbol{B}_{i,j}\right) \xi_{i,j} V_{i,j} \boldsymbol{B}_{i,j} = 0$. For example, if phase-three sampling is stratified on case status, we use weights $\tilde{w}_{i,j}^{(3)} = \frac{\sum_{l=1}^{J} \sum_{\substack{k=1, \\ non\,case}}^{n^{(j)}} \xi_{k,l}}{\sum_{l=1}^{J} \sum_{\substack{k=1, \\ non\,case}}^{n^{(j)}} \xi_{k,l} V_{k,l}}$ if subject $i$ in stratum $j$ is a non-case, and $\tilde{w}_{i,j}^{(3)} = \frac{\sum_{l=1}^{J} \sum_{\substack{k=1, \\ case}}^{n^{(j)}} \xi_{k,l}}{\sum_{l=1}^{J} \sum_{\substack{k=1, \\ case}}^{n^{(j)}} \xi_{k,l} V_{k,l}}$ if subject $i$ in stratum $j$ is a case. Finally, we estimate $\mathrm{var}\left(V_{i,j}\right) \equiv \sigma_{i,j}^{(3)}$ by $\widetilde{\sigma}_{i,j}^{(3)} = \frac{1}{\widetilde{w}_{i,j}^{(3)}} \left(1 - \frac{1}{\widetilde{w}_{i,j}^{(3)}}\right)$.

### 5.3 Estimation of relative hazard, cumulative baseline hazard and pure risk

We obtain the log-relative hazard estimate $\widetilde{\boldsymbol{\beta}}\left(\widetilde{\boldsymbol{\gamma}}\right) \equiv \widetilde{\boldsymbol{\beta}}$ from solving for $\boldsymbol{\beta}$ in the estimating equation $\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} \int_t V_{i,j} \widetilde{w}_{i,j}^{(3)} \left\{ \boldsymbol{X}_{i,j} - \frac{\widetilde{S}_1(t;\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta})}{\widetilde{S}_0(t;\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta})} \right\} \mathrm{d}N_{i,j}(t) = 0$. We let $\widetilde{w}_{k,j} = w_{i,j}^{(2)} \times \widetilde{w}_{i,j}^{(3)}$ and compute $\widetilde{S}_0\left(t;\widetilde{\boldsymbol{\gamma}}, \boldsymbol{\beta}\right)$, $\widetilde{\boldsymbol{S}}_1\left(t;\widetilde{\boldsymbol{\gamma}}, \boldsymbol{\beta}\right)$ and $\widetilde{\boldsymbol{S}}_2\left(t;\widetilde{\boldsymbol{\gamma}}, \boldsymbol{\beta}\right)$ from Eqs. (2)–(4) by substituting $\xi_{k,j} V_{k,j}$ for $\xi_{k,j}$ and $\widetilde{w}_{k,j}$ for $w_{k,j.}$ We estimate the baseline hazard point mass at time $t$, $\mathrm{d}\widetilde{\Lambda}_0\left(t;\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}}\right) \equiv \mathrm{d}\widetilde{\Lambda}_0(t)$, the cumulative baseline hazard up to time $t$, $\widetilde{\Lambda}_0\left(t;\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}}\right) \equiv \widetilde{\Lambda}_0(t)$, and the pure risk for profile $\boldsymbol{x}$ in the interval $(\tau_1, \tau_2]$, $\widetilde{\pi}\left(\tau_1, \tau_2; \boldsymbol{x}, \widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}}, \mathrm{d}\widetilde{\Lambda}_0\right) \equiv \widetilde{\pi}\left(\tau_1, \tau_2; \boldsymbol{x}\right)$, from Eqs. (5)–(7) with $\widetilde{S}_0\left(t;\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}}\right)$ and $\widetilde{\boldsymbol{\beta}}$ replacing $S_0\left(t, \widehat{\boldsymbol{\beta}}\right)$ and $\widehat{\boldsymbol{\beta}}$.

If a case with missing phase-two data occurs at a time $t$ when no other member of the phase-three sample is at risk, the Breslow estimate of cumulative baseline hazard is undefined. One option is to restrict the risk projection interval by increasing $\tau_1$ or decreasing $\tau_2$ to avoid such times. If there are only a small number of such times, we recommend ignoring them in all calculations.

### 5.4 Influence functions

Let $\boldsymbol{\Delta}_{i,j}\left(\widetilde{\boldsymbol{\theta}}\right)$ denote the influence of subject $i$ in stratum $j$ on one of the parameters $\widetilde{\boldsymbol{\theta}}$ from the set $\left\{\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}}, \mathrm{d}\widetilde{\Lambda}_0(t), \widetilde{\Lambda}_0(t), \widetilde{\pi}\left(\tau_1, \tau_2; \boldsymbol{x}\right)\right\}$, $i \in \left\{1, \dots, n^{(j)}\right\}$, $j \in \{1, \dots, J\}$. We can show that $\boldsymbol{\Delta}_{i,j}\left(\widetilde{\boldsymbol{\theta}}\right) = \xi_{i,j} \boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\theta}}\right) + \xi_{i,j} V_{i,j} \exp\left(\widetilde{\boldsymbol{\gamma}}' \boldsymbol{B}_{i,j}\right) \boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)$. Explicit forms of

$\boldsymbol{IF}_{i,j}^{(s)}\left(\widetilde{\boldsymbol{\theta}}\right)$, $s \in \{2, 3\}$, are given in Appendix 2 and are derived in Web Appendix E.1. The superscript 3 emphasizes that $\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)$ involves variables that are measured only on individuals in the phase-three sample. Thus $\xi_{i,j}V_{i,j}\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)$ is zero if individual $i$ in stratum $j$ is not in the phase-three sample. However, such an individual affects $\widetilde{\boldsymbol{\theta}}$ through her/his influence on $\widetilde{\boldsymbol{\gamma}}$ via $\xi_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\theta}}\right)$, as he/she is used to estimate the phase-three sampling weights.

### 5.5 Variance decomposition and estimation from influence functions

From $\operatorname{var}\left(\widetilde{\boldsymbol{\theta}}\right) \approx \operatorname{var}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\boldsymbol{\Delta}_{i,j}\left(\widetilde{\boldsymbol{\theta}}\right)\right\}$, $\widetilde{\boldsymbol{\theta}} \in \left\{\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}}, \mathrm{d}\widetilde{\Lambda}_0(t), \widetilde{\Lambda}_0(t), \widetilde{\pi}\left(\tau_1, \tau_2; \boldsymbol{x}\right)\right\}$, the variance $\operatorname{var}\left(\widetilde{\boldsymbol{\theta}}\right)$ can be decomposed as

$$\begin{aligned}&\operatorname{var}\left(\mathrm{E}\left[\mathrm{E}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\boldsymbol{\Delta}_{i,j}\left(\widetilde{\boldsymbol{\theta}}\right)|C_1, C_2\right\}|C_1\right]\right)\\&+\mathrm{E}\left(\operatorname{var}\left[\mathrm{E}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\boldsymbol{\Delta}_{i,j}\left(\widetilde{\boldsymbol{\theta}}\right)|C_1, C_2\right\}|C_1\right]\right)+\mathrm{E}\left(\mathrm{E}\left[\operatorname{var}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\boldsymbol{\Delta}_{i,j}\left(\widetilde{\boldsymbol{\theta}}\right)|C_1, C_2\right\}|C_1\right]\right),\end{aligned} \tag{21}$$

where $C_1$ denotes the information from the whole cohort, and $C_2$ denotes the information from the phase-two sample. The three terms correspond respectively to sampling from the "superpopulation", sampling the subcohort from the cohort, and sampling the phase-three sample from the phase-two sample.

We estimate $\operatorname{var}\left(\widetilde{\boldsymbol{\theta}}\right)$ by

$$\begin{aligned}&\frac{n}{n-1}\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\frac{1}{w_{i,j}^{(2)}}\left\{\xi_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\theta}}\right)' + 2\xi_{i,j}V_{i,j}\widetilde{w}_{i,j}^{(3)}\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)'\right.\\&\left.+\xi_{i,j}V_{i,j}\widetilde{w}_{i,j}^{(3)}\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)'\right\} + \sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\sigma_{i,j}^{(2)}w_{i,j}^{(2)}\left\{\xi_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\theta}}\right)'\right.\\&\left.+2\xi_{i,j}V_{i,j}\widetilde{w}_{i,j}^{(3)}\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)' + \xi_{i,j}V_{i,j}\widetilde{w}_{i,j}^{(3)}\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)'\right\}\\&+\sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\sum_{\substack{k=1\\k\neq i}}^{n^{(j)}}\sigma_{i,k,j}^{(2)}w_{i,k,j}^{(2)}\left\{\xi_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\theta}}\right) + \xi_{i,j}V_{i,j}\widetilde{w}_{i,j}^{(3)}\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)\right\}\left\{\xi_{k,j}\boldsymbol{IF}_{k,j}^{(2)}\left(\widetilde{\boldsymbol{\theta}}\right)\right.\\&\left.+\ \xi_{k,j}V_{k,j}\widetilde{w}_{k,j}^{(3)}\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)\right\}' + \sum_{j=1}^{J}\sum_{i=1}^{n^{(j)}}\widetilde{\sigma}_{i,j}^{(3)}\widetilde{w}_{i,j}^{(3)}\xi_{i,j}V_{i,j}\widetilde{w}_{i,j}^{(3)}\widetilde{w}_{i,j}^{(3)}\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\theta}}\right)'.\end{aligned} \tag{22}$$

See Web Appendix E.2 for details. Variability of the estimated phase-three weights is accounted for in a part of $\xi_{i,j}\boldsymbol{IF}_{i,j}^{(2)}\left(\tilde{\boldsymbol{\theta}}\right)$.

## 6 Software: CaseCohortCoxSurvival on CRAN

"Dfbetas", which approximate influences and are available from survival software, can be used to estimate the variance of $\widehat{\boldsymbol{\beta}}$ from unstratified case-cohort data (Therneau and Li 1999), and similar code was given to estimate the variance of $\widehat{\boldsymbol{\beta}}$ (which corresponds to Estimate II in Borgan et al. (2000)) for stratified case-cohort designs (Samuelsen et al. 2007). The cch function from the CRAN package survival (Therneau et al. 2023) deals with Estimate I and Estimate II of Borgan et al. (2000). The CRAN package cchs (Jones 2018, 2020) was created for Estimate III of Borgan et al. (2000), but we do not consider Estimate III. The twophase function from the CRAN package survey (Lumley 2021) estimates $\boldsymbol{\beta}$ and its variance from a phase-two sample, and thus from unstratified or stratified case-cohort data. The previous papers did not discuss pure risk, however. SAS code was proposed for Estimate III for stratified case-cohort studies and pure risk (Langholz and Jiao 2007), but we have been unable to find online procedures at the site mentioned in the original article. The CRAN package NestedCohort (Mark and Katki 2006) was removed from the CRAN repository, but its formerly available version can be found at https://dceg.cancer.gov/tools/analysis/nested-cohort and on the CRAN archive. More general survey software can accommodate weight calibration in addition to stratification (Lumley 2021), but R code showing how to use these more general programs to estimate $\boldsymbol{\beta}$, as referenced in Breslow et al. (2009a, 2009b), are no longer online. Thus, there is a need for convenient software to allow for stratification, weight calibration and missing phase-two data.

We have created the CaseCohortCoxSurvival CRAN package available at https://CRAN.R-project.org/package=CaseCohortCoxSurvival, to facilitate such analyses. We present a script in Table 4 to illustrate convenient analysis of mortality data from Golestan, Iran in Sect. 8. Extensive details on the features and arguments of the caseCohortCoxSurvival function will be provided elsewhere.

## 7 Simulations

### 7.1 Simulation designs

We compared how well the methods in Sects. 3.3 and 4.3 estimate the variance of the log-relative hazard and of pure risk estimates in simulated cohorts. We also evaluated the gain in precision from using calibrated weights rather than the design weights. We considered a range of scenarios, defined by the models described below and by parameter values in Web Tables 1 and 2 of Web Appendix D.1.

We simulated cohorts with $n \in \left\{5 \times 10^3, 10^4\right\}$ and used time on study as the time scale. We simulated three covariates $\boldsymbol{X} = \left(X_1, X_2, X_3\right)'$: $X_1 \sim \mathcal{N}(0,1)$, $X_2$ takes values in $\Omega_{X_2} = \{0,1,2\}$ with respective probabilities $\left\{p_{0|X_1}, p_{1|X_1}, p_{2|X_1}\right\}$, given in Web Table 1 in Web Appendix D.1, and $X_3 \sim \mathcal{N}\left(\alpha_1 X_1 + \alpha_2 X_2, 1\right)$, where $\mathcal{N}(a,b)$ denotes the Normal distribution with mean $a$ and variance $b$. We defined a categorical variable $W$ from $X_2$ and from a binary variable based on $X_1$: $W = 0 \times I\left(X_1 \geq 0, X_2 = 0\right) + 1 \times I\left(X_1 < 0, X_2 < 2\right) + 2 \times I\left(X_1 \geq 0, X_2 > 0\right)$ $+3 \times I\left(X_1 < 0, X_2 = 2\right)$, where $I()$ is the indicator function. We simulated proxies of $X_1$ and $X_3$, $\widetilde{X}_1$ and $\widetilde{X}_3$, as $\widetilde{X}_1 = X_1 + \varepsilon_1$, $\widetilde{X}_3 = X_3 + \varepsilon_3$, with $\varepsilon_1$ and $\varepsilon_3$ independently distributed as Normal $\mathcal{N}\left(0, 0.75^2\right)$, so that $\operatorname{corr}\left(\tilde{X}_1, X_1\right) = \operatorname{corr}\left(\tilde{X}_3, X_3\right) = 0.8$. We simulated failure time $T$ from a Cox proportional hazards model with hazard $\lambda(t;\boldsymbol{X}) = \lambda_0 \times \exp\left(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\right)$, where the baseline hazard $\lambda_0 = \frac{p_Y}{\mathrm{E}\left\{\exp\left(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\right)\right\} \times 10}$, $p_Y \in \{0.02, 0.05, 0.1\}$, is a constant calculated to have approximately 98%, 95% or 90% 10-year pure survival probability. Parameters $\alpha_1$, $\alpha_2$, $\beta_1$, $\beta_2$ and $\beta_3$ are in Web Table 2 in Web Appendix D.1. Cohort entry time, $E$, was uniform on the first 5 years, and we assumed the time to censoring by loss to follow-up, $C$, had an exponential distribution with hazard $\frac{10}{-\log(0.98)}$, corresponding to a pure risk of loss to follow-up of 2% in 10 years. We assumed $T$, $E$ and $C$ were mutually independent. We let $\widetilde{T} = \min(T, 10 - E, C)$ be the observed time. The total follow-up time on time interval $(\tau_1, \tau_2]$ was thus $\max\left\{0, \min\left(\tilde{T}, \tau_2\right) - \tau_1\right\}$.

We assumed that $X_2$, $W$, $\widetilde{X}_1$, $\widetilde{X}_3$, $\widetilde{T}$ and the case status were known for everybody in the cohort, but $X_1$ and $X_3$ were available only for individuals in phase-two. We sampled from the four strata defined by $W$; thus stratum "0" is low risk, strata "1" and "2" are both medium risk, and stratum "3" is high risk. We sampled without replacement fixed numbers of individuals, $m^{(j)}$, independently across strata, $j \in \{0,1,2,3\}$, and independently of the case status. The $m^{(j)}$ depended on the expected numbers of failures and of individuals in the strata via $m^{(j)} = \left\lfloor \frac{\lambda_0 \times 10 \times \mathrm{E}\left\{\exp\left(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\right) | W = j\right\}}{1 - \lambda_0 \times 10 \times \mathrm{E}\left\{\exp\left(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\right) | W = j\right\}} \times \mathrm{E}\left(n^{(j)}\right) \times K + \frac{1}{2} \right\rfloor$, where $K \in \{2,4\}$ is the number of non-cases we wish to sample for each case, and $\lfloor \quad \rfloor$ is the floor function. The subcohort consisted of all the sampled subjects from the $J$ strata. Then, we sample all the cases in the cohort (some may have been included in the subcohort); the phase-two (or case-cohort) sample consisted of the subcohort and all the cases. Design weights were computed as in Sect. 3.3. Calibration of the weights was performed against the auxiliary variables proposed in Sect. 4.1, with the values of covariates $X_1$ and $X_3$ in the full cohort imputed from weighted linear regressions of $X_1$ on $\widetilde{X}_1$ and $W$, and of $X_3$ on $\widetilde{X}_1$ and $\widetilde{X}_3$.

For each scenario, we simulated 5000 cohorts. We estimated the log-relative hazard $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \beta_3\right)'$ and pure risks $\pi\left(\tau_1, \tau_2; \boldsymbol{x}\right)$ in time interval $(\tau_1, \tau_2] = (0, 8]$ and for covariate profiles $\boldsymbol{x} \in \left\{(-1, 1, -0.6)', (1, -1, 0.6)', (1, 1, 0.6)'\right\}$, using the following sampling designs and methods of analysis: the stratified case-cohort with design weights (SCC); the stratified case-cohort with calibrated weights (SCC.Calib); the unstratified case-cohort with design weights (USCC); and the unstratified case-cohort

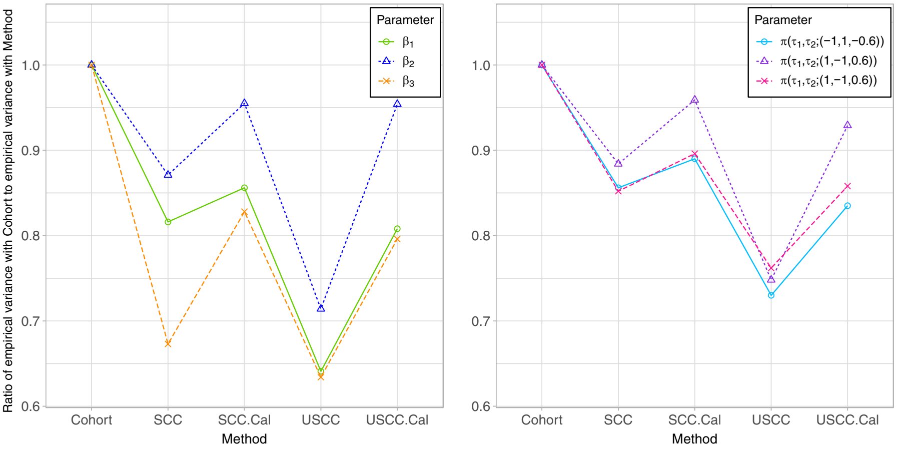

**Fig. 1** Ratio of empirical variance of log-relative hazard and pure risk estimates with the whole cohort to that when using different sampling designs and methods of analysis. The results are obtained from 5000 simulated cohorts with $n = 10{,}000$, $p_Y = 0.02$, $K = 2$. The variance ratio is a measure of relative efficiency

**Table 1** Mean of estimated variances of log-relative hazard and pure risk estimates, from using different sampling designs, methods of analysis and variance estimation, over 5000 simulated cohorts with $n = 10{,}000$, $p_Y = 0.02$, $K = 2$

| Parameter | Cohort | SCC | | SCC.Calib | | USCC | | USCC.Calib | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ |
| $\boldsymbol{\beta}$ | | | | | | | | | |
| $\beta_1$ | 0.0069 | 0.0102 | 0.0087 | 0.0084 | 0.0082 | 0.0106 | 0.0106 | 0.0086 | 0.0086 |
| | (0.007) | (0.0085) | | (0.0081) | | (0.0108) | | (0.0087) | |
| $\beta_2$ | 0.0097 | 0.0139 | 0.0114 | 0.0103 | 0.0102 | 0.014 | 0.014 | 0.0103 | 0.0103 |
| | (0.01) | (0.0115) | | (0.0105) | | (0.014) | | (0.0105) | |
| $\beta_3$ | 0.0068 | 0.0102 | 0.0102 | 0.0084 | 0.0084 | 0.0107 | 0.0107 | 0.0087 | 0.0087 |
| | (0.0069) | (0.0103) | | (0.0083) | | (0.0109) | | (0.0087) | |
| $\log\{\pi(\tau_1,\tau_2;\boldsymbol{x})\}$ | | | | | | | | | |
| $\boldsymbol{x} = (-1, 1, -0.6)'$ | 0.0122 | 0.0172 | 0.014 | 0.0142 | 0.0137 | 0.0181 | 0.0159 | 0.0145 | 0.0145 |
| | (0.0119) | (0.0136) | | (0.0133) | | (0.0158) | | (0.014) | |
| $\boldsymbol{x} = (1, -1, 0.6)'$ | 0.062 | 0.0861 | 0.0697 | 0.0664 | 0.066 | 0.086 | 0.0837 | 0.0676 | 0.0676 |
| | (0.0618) | (0.0688) | | (0.0649) | | (0.0823) | | (0.0676) | |
| $\boldsymbol{x} = (1, 1, 0.6)'$ | 0.0277 | 0.0379 | 0.0333 | 0.0318 | 0.0315 | 0.0386 | 0.0363 | 0.0327 | 0.0327 |
| | (0.0274) | (0.0326) | | (0.0308) | | (0.0361) | | (0.0323) | |

The corresponding empirical variances are displayed between parentheses

with calibrated weights (USCC.Calib). We then estimated their variance. For each simulated realization, we obtained the variance estimate $\widehat{V}$ for SCC from Eq. (14) and the robust variance estimate ($\widehat{V}_{\text{Robust}}$) from Eq. (15). For SCC.Calib, we used $\widehat{V}$ in Eq. (18) and $\widehat{V}_{\text{Robust}}$ in Eq. (19). For USCC and USCC.Calib, we used the variance

**Table 2** Coverage of 95% CIs of log-relative hazard and pure risk estimates, from using different sampling designs, methods of analysis and variance estimation, over 5,000 simulated cohorts with $n =$ 10,000, $p_Y = 0.02$, $K = 2$

| Parameter | Cohort | SCC | | SCC.Calib | | USCC | | USCC.Calib | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ |
| $\boldsymbol{\beta}$ | | | | | | | | | |
| $\beta_1$ | 0.9476 | 0.9668* | 0.9524 | 0.953 | 0.9506 | 0.947 | 0.947 | 0.9464 | 0.9464 |
| $\beta_2$ | 0.9486 | 0.9716* | 0.9522 | 0.9498 | 0.9496 | 0.9554 | 0.9554 | 0.9466 | 0.9466 |
| $\beta_3$ | 0.9422* | 0.9492 | 0.9494 | 0.9498 | 0.9498 | 0.9454 | 0.9454 | 0.9482 | 0.9482 |
| $\log\{\pi(\tau_1,\tau_2;\boldsymbol{x})\}$ | | | | | | | | | |
| $\boldsymbol{x} = (-1, 1, -0.6)'$ | 0.956* | 0.973* | 0.9568* | 0.9604* | 0.9554 | 0.9656* | 0.9526 | 0.956* | 0.956* |
| $\boldsymbol{x} = (1, -1, 0.6)'$ | 0.952 | 0.9722* | 0.9518 | 0.9542 | 0.9532 | 0.954 | 0.9522 | 0.95 | 0.95 |
| $\boldsymbol{x} = (1, 1, 0.6)'$ | 0.952 | 0.9666* | 0.9542 | 0.9538 | 0.952 | 0.9574* | 0.95 | 0.9502 | 0.9502 |

*Indicates coverage outside the expected interval [0.9440; 0.9560]

estimates in Eqs. (14), (15), (18) and (19) with $J = 1$. Corresponding 95% confidence intervals (CIs) were computed assuming normality. As a point of reference, we also estimated these parameters using the data from the whole cohort (Cohort).

## 7.2 Simulation results

The simulation results for the scenario with $n = 10{,}000$, $p_Y = 0.02$ and $K = 2$ are displayed in Fig. 1, Tables 1 and 2; see Web Table 3 to 20 in Web Appendix D.2 for other scenarios. The robust variance formula overestimated the variance (Table 1) and yielded supra-nominal confidence interval coverage (Table 2) for most log-relative hazards and pure risks with stratified designs, and for pure risk with unstratified designs. Weight calibration led to smaller variances (Table 1), as expected, because it led to a smaller phase-two component of the variance (see also Web Appendix D.4). In addition, robust variance estimates were approximately valid with calibrated weights, except for one pure risk (Tables 1 and 2). Because they properly accounted for the sampling features, the variance estimates in Eq. (14) for design weights and Eq. (18) for calibrated weights yielded proper coverage in all designs (Table 2), except for $\log\left\{\pi\left(\tau_1, \tau_2; \boldsymbol{x}\right)\right\}$ when $\boldsymbol{x} = (-1, 1, -0.6)'$, for which the full cohort analysis also had supra-nominal coverage. As shown in Fig. 1, stratification and/or weight calibration improved efficiency. Moreover, the unstratified case-cohort with weight calibration was nearly as efficient as the stratified case-cohort with weight calibration, and both were considerably more efficient than analyses with design weights. With design weights, stratification improved efficiency compared to the unstratified case-cohort design.

A few remarks follow. First, variables $X_1$ and $X_3$ were only measured in the phase-two sample. The strongest increase in efficiency from calibration was usually for $\widehat{\beta}_2$, because $X_2$ was measured in the entire cohort (see also Sect. 8). With weaker proxies, the efficiency gain from calibration would be more modest, and robust variance estimates may be too large (see Web Appendix D.6). Second, for log-relative hazards, the nominal coverage of the 95% CIs suggested that inference can be based on asymptotic normality, even with calibrated weights. We log-transformed the pure risks to improve coverage based on asymptotic Normal theory. Third, some authors used *post-stratified* weights instead of design weights, by having a separate stratum for cases and excluding cases from the strata with non-cases (Borgan et al. 2000; Samuelsen et al. 2007). This approach improved the precision of estimates with SCC and USCC negligibly (variance ratios of 1.01 or less), compared to using design weights (Web Appendix D.5). Finally, each stratum in the cohort and in the case-cohort had substantial numbers of subjects. Unreported simulation with very few subjects and cases in stratum $W = 0$ led to similar results.

Simulations concerning missing phase-two data showed that Eq. (22) in Sect. 5.5 and a simpler formula that ignores variability in the estimated weights (Web Appendix F.3) yielded nominal confidence interval coverage of log-relative hazards and pure risk (Web Appendix G), but in non-reported simulations with larger proportions missing, the simpler formula overestimated the variance. We therefore recommend using Eq. (22), as is computed in CaseCohortCoxSurvival available at https://CRAN.R-project.org/package=CaseCohortCoxSurvival.

## 8 Data analysis

The Golestan Cohort included 49,819 individuals aged 36–81 and recruited in 2003–2009 (Pourshams et al. 2010). To reduce computation, we randomly sampled $n=30{,}000$ individuals and analyzed this subset. We used the age-scale and assumed a Cox proportional hazards model predicting mortality from baseline variables: $X_1$=indicator of male gender, $X_2$=wealth score, $X_3$=indicator of former smoker (cigarettes, nass, or opium), $X_4$=indicator of current smoker, $X_5$=indicator of morbidity, $X_6 = X_1 X_3$ and $X_7 = X_1 X_4$. We used “never smoker” as the reference category, and morbidity was a binary indicator with value 1 if the individuals had at least one of the following morbidities at baseline: cardiovascular disease, cerebrovascular accident, hypertension, diabetes, chronic obstructive pulmonary disease, tuberculosis, cancer. The wealth score had been computed from information such as house ownership and number and type of household appliances; see Islami et al. (2009). We also estimated the pure risk in interval $(\tau_1, \tau_2] = (52, 66]$ and for covariate profiles $\boldsymbol{x} \in \left\{ \left(0, -0.4, 0, 1, \mathbf{0}_3\right)', \left(0, 0.4, 0, 1, \mathbf{0}_3\right)', \left(\mathbf{0}_4, 1, \mathbf{0}_2\right)', \mathbf{0}_7' \right\}$, where $\mathbf{0}_a$ is the $a \times 1$ vector of zeros, and where for example $\left(0, -0.4, 0, 1, \mathbf{0}_3\right)'$ corresponds to the profile of a currently smoking woman with a low wealth score, while $\left(\mathbf{0}_4, 1, \mathbf{0}_2\right)'$ corresponds to a never-smoking woman with morbidity at baseline. We assumed that age, gender, smoking status, morbidity, residence (urban, rural), ethnicity (Turkmen, others), marital status (unmarried, married, widowed, divorced/separated, other), education (nil, less than 5th, 6th-8th, 9th-12th, College), socioeconomic status (low, low to medium, medium to high, high), death status and follow-up time were known for everybody in the cohort, but the wealth score was available only for individuals in phase-two. We sampled 33, 42, 192,

**Table 3** Estimated variances of log-relative hazard and pure risk parameters from using different sampling designs, methods of analysis and variance estimation, in the Golestan Cohort ($n = 30{,}000$)

| Parameter | Cohort | SCC | | SCC.Calib | | USCC | | USCC.Calib | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ | $\widehat{V}_{\text{Robust}}$ | $\widehat{V}$ |
| $\boldsymbol{\beta}$ | | | | | | | | | |
| $\beta_1$ | 0.0013 | 0.0018 | 0.0016 | 0.0013 | 0.0013 | 0.0025 | 0.0025 | 0.0013 | 0.0013 |
| $\beta_2$ | 0.005 | 0.0075 | 0.0072 | 0.0055 | 0.0055 | 0.0104 | 0.0104 | 0.0061 | 0.0061 |
| $\beta_3$ | 0.0173 | 0.0289 | 0.0289 | 0.0178 | 0.0178 | 0.0394 | 0.0394 | 0.0173 | 0.0173 |
| $\beta_4$ | 0.0029 | 0.0057 | 0.0057 | 0.0029 | 0.0029 | 0.007 | 0.007 | 0.0031 | 0.0031 |
| $\beta_5$ | 0.0009 | 0.0013 | 0.0013 | 0.0009 | 0.0009 | 0.002 | 0.002 | 0.0009 | 0.0009 |
| $\beta_6$ | 0.0217 | 0.0342 | 0.0342 | 0.0219 | 0.0219 | 0.0497 | 0.0497 | 0.0217 | 0.0217 |
| $\beta_7$ | 0.0044 | 0.0076 | 0.0076 | 0.0044 | 0.0044 | 0.0103 | 0.0103 | 0.0045 | 0.0045 |
| $\pi(\tau_1, \tau_2; \boldsymbol{x})$ | | | | | | | | | |
| $\boldsymbol{x} = (0, -0.4, 0, 1, \mathbf{0}_3)'$ | 0.00012 | 0.00023 | 0.00022 | 0.00012 | 0.00012 | 0.00031 | 0.00030 | 0.00013 | 0.00013 |
| $\boldsymbol{x} = (0, 0.4, 0, 1, \mathbf{0}_3)'$ | 5.10E−05 | 9.20E−05 | 8.90E−05 | 5.20E−05 | 5.10E−05 | 0.00010 | 0.00010 | 5.70E−05 | 5.70E−05 |
| $\boldsymbol{x} = (\mathbf{0}_4, 1, \mathbf{0}_2)'$ | 2.50E−05 | 4.10E−05 | 3.50E−05 | 2.50E−05 | 2.50E−05 | 4.50E−05 | 4.30E−05 | 2.70E−05 | 2.70E−05 |
| $\boldsymbol{x} = \mathbf{0}_7'$ | 5.60E−06 | 8.50E−06 | 6.70E−06 | 5.80E−06 | 5.70E−06 | 8.90E−06 | 8.30E−06 | 5.60E−06 | 5.60E−06 |

$\beta_p$ denotes the log-relative hazard parameter of covariate $X_p$, $p \in \{1, \ldots, 7\}$

**Table 4** R script using the CaseCohortCoxSurvival R package to obtain variance estimates $\widehat{V}$ and $\widehat{V}_{\text{Robust}}$ with SCC and SCC.Calib, for the log-relative hazards and pure risk with profile $\boldsymbol{x} = \left(0, -0.4, 0, 1, \mathbf{0}_3\right)'$ in Table 3

```
library(CaseCohortCoxSurvival)


# Loading the data set --------------------------------------------------

load("Golestancohort.RData")


# Estimation using the stratified case-cohort with design weights -----------

caseCohortCoxSurvival(data = Golestancohort, status = "status", time =
c(“agebegin”, “ageend”), cox.phase1 = c(“x1”, “x3”, “x4”, “x5”, “x6”, “x7”),
cox.phase2 = “x2”, strata = "gender.residence.age", subcohort = "subcohort",
Tau1 = 52, Tau2 = 66, x = list(x1 = 0, x2 = -0.4, x3 = 0, x4 = 1, x5 = 0,
x6 = 0, x7 = 0))



# Estimation using the stratified case-cohort with calibrated weights --------

caseCohortCoxSurvival(data = Golestancohort, status = "status", time =
c(“agebegin”, “ageend”), cox.phase1 = c(“x1”, “x3”, “x4”, “x5”, “x6”, “x7”),
cox.phase2 = “x2”, strata = "gender.residence.age", subcohort = "subcohort",
Tau1 = 52, Tau2 = 66, x = list(x1 = 0, x2 = -0.4, x3 = 0, x4 = 1, x5 = 0,
x6 = 0, x7 = 0), calibrated = TRUE, predictors.cox.phase2 = list(x2 = c(“x1”,
“ses”, “age”, “maritalstatus”, “ethnicity”, “education”, “residence”)),
aux.method = “Shin”)
```

246, 57, 62, 313, 382, 82, 86, 391, 477, 565, 770, 1934 and 2949 individuals respectively in the 16 strata defined by gender (male, female), residence and four baseline age categories ([36,45), [45,50), [50,55) and [55,81]), so that we expected approximately one non-case per case in each stratum. We estimated the log-relative hazards and pure risks using SCC, SCC.Calib, USCC and USCC.Calib (see notation and methods of analysis in Sect. 5). We used gender, socioeconomic status, age at baseline, marital status, ethnicity, education and residence as proxies to impute the wealth score for the entire cohort and then calibrated the design weights. We also analyzed the whole cohort ($n$=30,000).

Table 3 displays the variances of log-relative hazard and pure risk parameters; see Web Table 55 in Web Appendix H for parameter estimates. When using design weights, robust variance estimates were larger for the log-relative hazards of covariates $X_1$ and $X_2$, for all the pure risks in the stratified design, and for the pure risks with profiles $\boldsymbol{x} \in \left\{ \left(0, -0.4, 0, 1, \mathbf{0}_3\right)', \left(\mathbf{0}_4, 1, \mathbf{0}_2\right)', \mathbf{0}_7' \right\}$ in the unstratified design. In the stratified design, $\widehat{V}_{\text{Robust}}$ agreed well with $\widehat{V}$ for 5 of the 7 log-relative hazard parameters, possibly because stratification was only based on $X_1$.

Weight calibration improved efficiency, and robust variance estimates were very close to $\widehat{V}$ for all parameters. Notably, calibration led to estimates with almost as much precision as with the full cohort, not only for covariates that were available on the whole cohort, but also for wealth score, for which there were good proxies.

Web Appendix I presents pseudo-code for all of the steps for estimation in Sects. 3 and 4. To illustrate how easily such analyses can be performed with the CaseCohortCoxSurvival CRAN package (available at https://CRAN.R-project.org/package=CaseCohortCoxSurvival), we present a script for SCC and SCC.Calib in Table 4.

## 9 Discussion

We presented a unified approach to analysis of case-cohort data that allows the practitioner to take advantage of various options and improvements in design and analysis since the landmark paper of Prentice (1986). We used influence functions adapted to the various design and analysis options together with variance calculations that take two-phase sampling into account. We developed corresponding software CaseCohortCoxSurvival, available at https://CRAN.R-project.org/package=CaseCohortCoxSurvival, that facilitates analysis with and without stratification and/or weight calibration, for subcohort sampling with or without replacement. We allow for phase-two data to be missing at random for stratified designs. We provide inference not only for log-relative hazards in the Cox model, but also for covariate-specific cumulative hazards and pure risks. We hope these calculations and software will promote wider and more principled design and analysis of case-cohort data, for which there is a need (Sharp et al. 2014). Detailed features and arguments of the CaseCohortCoxSurvival CRAN R package available at https://CRAN.R-project.org/package=CaseCohortCoxSurvival will be described elsewhere. Convenient software of the type we propose does not appear to be available online (Sect. 6).

We found that weight calibration improves efficiency with stratified or unstratified sampling of the subcohort, in line with previous findings for unstratified designs. We found theoretically and empirically that the robust variance estimate (Barlow 1994) is nearly unbiased if the covariances of the phase-two sampling indicators, $\sigma_{i,k,j}$, $i \neq k$, are zero, as when the subcohort members are sampled with replacement (Table 5). For sampling without replacement, these covariances are negative, which tends to bias the robust variance estimate upward. This has been noted for log-relative hazards in stratified designs (Gray 2009; Samuelsen et al. 2007), but we also found this bias for pure risk in unstratified designs. With weight calibration based on strong predictors of phase-two covariates, the robust variance had little bias (Table 5). Nonetheless, we recommend our influence-based approach with complete variance decomposition for theoretical and empirical reasons. In addition, and as previously recommended (Sharp et al. 2014), we stress the practical importance of describing the design fully in publications, including stratification details and whether or not the subcohort was sampled with replacement.

**Table 5** Sampling designs and methods of analysis for which the robust variance estimator is approximately valid for relative hazard and pure risk estimates

| Parameter | Sampling with replacement | Sampling without replacement | | | |
|---|---|---|---|---|---|
| | | Unstratified sampling with design weights | Unstratified sampling with calibrated weights | Stratified sampling with design weights | Stratified sampling with calibrated weights* |
| *Approximate validity of robust variance estimate* | | | | | |
| Relative hazard | Yes | Yes | Yes | No | Yes* |
| Pure risk | Yes | No | Yes* | No | Yes* |

We categorize the robust variance estimate as approximately valid if theoretical calculations indicate that the bias is negligible and/or if in simulations the means of the robust variance estimates were close to the means of the two-phase variance procedures we describe.

*Holds if the phase-one covariates are good predictors of covariates measured only in phase-two. If the proxies of phase-two covariates are too weak, the robust variance estimate may be inappropriate

In our simulation, the unstratified case-cohort with calibrated weights was nearly as efficient as the stratified case-cohort with calibrated weights. This is probably because information used to define the strata was also used for calibration. However, if strata depended for example only on time of events or censoring, but imputation of phase-two covariates depended on other phase-one covariates, a stratified calibrated approach might be more efficient than the corresponding unstratified calibrated one.

To obtain the subcohort, we sampled fixed numbers of individuals from the strata, independently of the case status. Thus, some of the cases may have been included in the subcohort. Alternatively, one might want to sample fixed numbers of non-cases. To do so, the strata could be redefined by excluding the cases, and an additional stratum containing all the cases, could be created. It is possible to analyze data from this slightly modified design CRAN R package (available at https://CRAN.R-project.org/package=CaseCohortCoxSurvival) by sampling all the cases from the case stratum, so that all cases are included in the case-cohort and have unit design weights.

We focused on the "standard" case-cohort design, where all the cases are sampled from the cohort. If the event of interest is not rare, one may want to only include a fraction of the cases in the case-cohort. The derivations presented in this paper can be extended to such a design, sometimes called the *generalized case-cohort* design (Kim et al. 2018; Xu et al. 2022). For example, a weighted version $\xi_{i,j} w_{i,j} \mathrm{d}N_{i,j}(t)$ would be employed in Eqs. (1) and (5), with the cases having non-unit design weights. The cases would then contribute to the phase-two component of the variance.

The methods we presented used design weights. Borgan et al. (2000) and Samuelsen et al. (2007) recommended weights that are post-stratified into a case stratum and multiple non-case strata. In our simulations, there was less than 2% increase in efficiency from post-stratification. Using the influences we derived for design weights with post-stratified weights (with cases in one stratum and non-cases in the original strata) yielded confidence intervals with nominal coverage (results not shown). Thus, the influence functions we provide can be used for such post-stratification. Further efficiency gains might be obtained by post-stratifying on time intervals in which follow-up ends (Chen 2001; Ding et al. 2017; Samuelsen et al. 2007) or on other features (Section 16.4.5 in Borgan et al. 2017).

An alternative approach to sampling is to select the subcohort sample size in each stratum such that the expected number of non-cases is a multiple of the observed number of cases. In unreported simulations, using the influences we gave for design weights and substituting post-stratified weights yielded valid variance estimates and coverage of confidence intervals, unless the number of cases and non-cases in a stratum is small (e.g. fewer than 10 cases and 20 non-cases).

As discussed by Keogh et al. (2018), likelihood-based methods for missing data and imputation can increase efficiency of case-cohort analyses, but, unlike stratification and weight calibration, they yield biased risk model estimates if imputation models are misspecified. Indeed, a key advantage of weight calibration is that poor imputation models reduce the efficiency gains, but do not bias estimates of risk model parameters (Lumley et al. 2011). Weight calibrated estimators are in the class of augmented inverse-probability weighted estimators that are similarly robust (Lumley et al. 2011; Robins et al. 1994).

In Sect. 5.3, we suggested modifications for times when a case had missing covariate data and no other member of the phase-three sample was at risk when the case failed. An alternative would be to weight the numerator of the Breslow estimator and only use event times $t$ from cases with complete covariate data, namely $\mathrm{d}\widetilde{\Lambda}_0(t) = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n^{(j)}} V_{i,j} \widetilde{w}_{i,j}^{(3)} \mathrm{d}N_{i,j}(t)}{\widetilde{S}_0\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)}$. Unreported simulations showed this led to biased estimates of pure risks, however.

This paper dealt with covariates measured at baseline. Although the influences for log-relative hazards apply equally to time-varying covariates, modifications are needed for pure risks, and computational challenges arise for large cohorts. Moreover, pure risk estimates are uninterpretable unless the time-varying covariates are "external" (Kalbfleisch and Prentice 2011). We have assumed a common baseline hazard across strata. A stratified Cox model with different baseline hazards in each stratum would require modifications of the influences given in this paper.

## Appendix 1: Influences for stratified case-cohort with calibrated weights

$$\boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\eta}}\right) = \left\{ \sum_{l=1}^{J} \sum_{k=1}^{n^{(j)}} \xi_{k,l} w_{k,l} \exp\left(\widehat{\boldsymbol{\eta}}' \boldsymbol{A}_{k,l}\right) \boldsymbol{A}_{k,l} \boldsymbol{A}_{k,l}' \right\}^{-1} \boldsymbol{A}_{i,j},$$

$$\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\eta}}\right) = -\exp\left(\widehat{\boldsymbol{\eta}}' \boldsymbol{A}_{i,j}\right) \left\{ \sum_{l=1}^{J} \sum_{k=1}^{n^{(j)}} \xi_{k,l} w_{k,l} \exp\left(\widehat{\boldsymbol{\eta}}' \boldsymbol{A}_{k,l}\right) \boldsymbol{A}_{k,l} \boldsymbol{A}_{k,l}' \right\}^{-1} \boldsymbol{A}_{i,j},$$

$$\boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\beta}}^{*}\right) = \left\{ \sum_{l=1}^{J} \sum_{k=1}^{n^{(j)}} \xi_{k,l} w_{k,l} \exp\left(\widehat{\boldsymbol{\eta}}' \boldsymbol{A}_{k,l}\right) \boldsymbol{Z}_{k,l} \boldsymbol{A}_{k,l}' \right\} \boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\eta}}\right),$$

and $$\boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\beta}}^{*}\right) = \exp\left(\widehat{\boldsymbol{\eta}}' \boldsymbol{A}_{i,j}\right) \boldsymbol{Z}_{i,j} + \left\{ \sum_{l=1}^{J} \sum_{k=1}^{n^{(j)}} \xi_{k,l} w_{k,l} \exp\left(\widehat{\boldsymbol{\eta}}' \boldsymbol{A}_{k,l}\right) \boldsymbol{Z}_{k,l} \boldsymbol{A}_{k,l}' \right\} \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\eta}}\right),$$

with $$\boldsymbol{Z}_{i,j} = \left[ \sum_{l=1}^{J} \sum_{k=1}^{n^{(l)}} \int_t \xi_{k,l} w_{k,l} \exp\left(\widehat{\boldsymbol{\eta}}' \boldsymbol{A}_{k,l}\right) \left\{ \frac{\boldsymbol{S}_2^{*}\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^{*}\right)}{S_0^{*}\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^{*}\right)} - \frac{\boldsymbol{S}_1^{*}\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^{*}\right) \boldsymbol{S}_1^{*}\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^{*}\right)'}{S_0^{*}\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^{*}\right)^2} \right\} \mathrm{d}N_{k,l}(t) \right]^{-1}$$

$$\left[ \int_t \left\{ \boldsymbol{X}_{i,j} - \frac{\boldsymbol{S}_1^{*}\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^{*}\right)}{S_0^{*}\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^{*}\right)} \right\} \left\{ \mathrm{d}N_{i,j}(t) - \frac{Y_{i,j}(t) \exp\left(\widehat{\boldsymbol{\beta}}^{*\prime} \boldsymbol{X}_{i,j}\right) \sum_{l=1}^{J} \sum_{k=1}^{n^{(l)}} \xi_{k,l} w_{k,l} \exp\left(\widehat{\boldsymbol{\eta}}' \boldsymbol{A}_{k,l}\right) \mathrm{d}N_{k,l}(t)}{S_0^{*}\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^{*}\right)} \right\} \right].$$

$$IF_{i,j}^{(1)}\left\{\mathrm{d}\widehat{\Lambda}_0^*(t)\right\} = \left\{\sum_{l=1}^{J}\sum_{k=1}^{n^{(l)}} \xi_{k,l} w_{k,l} \exp\left(\widehat{\boldsymbol{\eta}}'\boldsymbol{A}_{k,l}\right) H_{k,l}(t)\boldsymbol{A}_{k,l}'\right\} \boldsymbol{IF}_{i,j}^{(1)}\left(\widehat{\boldsymbol{\eta}}\right),$$

$$IF_{i,j}^{(2)}\left\{\mathrm{d}\widehat{\Lambda}_0^*(t)\right\} = \left\{S_0^*\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^*\right)\right\}^{-1}\Bigg[\mathrm{d}N_{i,j}(t) + \exp\left(\widehat{\boldsymbol{\eta}}'\boldsymbol{A}_{i,j}\right)H_{i,j}(t) + \left\{\sum_{l=1}^{J}\sum_{k=1}^{n^{(l)}} \xi_{k,l} w_{k,l} \exp\left(\widehat{\boldsymbol{\eta}}'\boldsymbol{A}_{k,l}\right) H_{k,l}(t)\boldsymbol{A}_{k,l}'\right\} \boldsymbol{IF}_{i,j}^{(2)}\left(\widehat{\boldsymbol{\eta}}\right)\Bigg],$$

with $H_{i,j}(t) = -\left\{S_0^*\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^*\right)\right\}^{-1}\mathrm{d}\widehat{\Lambda}_0^*(t)\left\{\boldsymbol{S}_1^*\left(t;\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\beta}}^*\right)'\boldsymbol{Z}_{i,j} + K_{i,j}(t)\right\}$, and $K_{i,j}(t) = Y_{i,j}(t)\exp\left(\widehat{\boldsymbol{\beta}}^{*\prime}\boldsymbol{X}_{i,j}\right)$.

Finally, for any $s \in \{1,2\}$, $IF_{i,j}^{(s)}\left\{\int_{\tau_1}^{\tau_2}\mathrm{d}\widehat{\Lambda}_0^*(t)\right\} = \int_{\tau_1}^{\tau_2} IF_{i,j}^{(s)}\left\{\mathrm{d}\widehat{\Lambda}_0^*(t)\right\}$, and

$$IF_{i,j}^{(s)}\left\{\widehat{\pi}^*\left(\tau_1,\tau_2;\boldsymbol{x}\right)\right\} = \left\{\frac{\partial\widehat{\pi}\left(\tau_1,\tau_2;\boldsymbol{x}\right)}{\partial\boldsymbol{\beta}}_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^*}\right\}\boldsymbol{IF}_{i,j}^{(s)}\left(\widehat{\boldsymbol{\beta}}^*\right) + \left[\frac{\partial\widehat{\pi}\left(\tau_1,\tau_2;\boldsymbol{x}\right)}{\partial\left\{\int_{\tau_1}^{\tau_2}\mathrm{d}\Lambda_0(t)\right\}}_{|\mathrm{d}\Lambda_0(t)=\mathrm{d}\widehat{\Lambda}_0^*(t)}\right] IF_{i,j}^{(s)}\left\{\int_{\tau_1}^{\tau_2}\mathrm{d}\widehat{\Lambda}_0^*(t)\right\}.$$

## Appendix 2: Influences for stratified case-cohort with missing covariate information and estimated design sampling phase-three weights

$\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\gamma}}\right) = \left\{\sum_{l=1}^{J}\sum_{k=1}^{n^{(l)}}\xi_{k,l}V_{k,l}\exp\left(\widetilde{\boldsymbol{\gamma}}'\boldsymbol{B}_{k,l}\right)\boldsymbol{B}_{k,l}\boldsymbol{B}_{k,l}'\right\}^{-1}\boldsymbol{B}_{i,j}$, and

$\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\gamma}}\right) = -\left\{\sum_{l=1}^{J}\sum_{k=1}^{n^{(l)}}\xi_{k,l}V_{k,l}\exp\left(\widetilde{\boldsymbol{\gamma}}'\boldsymbol{B}_{k,l}\right)\boldsymbol{B}_{k,l}\boldsymbol{B}_{k,l}'\right\}\boldsymbol{B}_{i,j}$.

Then $\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\beta}}\right) = \left\{\sum_{l=1}^{J}\sum_{k=1}^{n^{(l)}}\xi_{k,l}V_{k,l}\exp\left(\widetilde{\boldsymbol{\gamma}}'\boldsymbol{B}_{k,l}\right)\widetilde{\mathbf{Z}}_{k,l}\boldsymbol{B}_{k,l}'\right\}\boldsymbol{IF}_{i,j}^{(2)}\left(\widetilde{\boldsymbol{\gamma}}\right)$, and $\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\beta}}\right) = \widetilde{\mathbf{Z}}_{i,j} + \left\{\sum_{l=1}^{J}\sum_{k=1}^{n^{(l)}}\xi_{k,l}V_{k,l}\exp\left(\widetilde{\boldsymbol{\gamma}}'\boldsymbol{B}_{k,l}\right)\widetilde{\mathbf{Z}}_{k,l}\boldsymbol{B}_{k,l}'\right\}\boldsymbol{IF}_{i,j}^{(3)}\left(\widetilde{\boldsymbol{\gamma}}\right)$, with

$$\widetilde{\mathbf{Z}}_{i,j} = w_{i,j}^{(2)} \times \left[\sum_{l=1}^{J}\sum_{k=1}^{n^{(l)}}\int_t \xi_{k,l}V_{k,l}w_{k,l}^{(2)}\exp\left(\widetilde{\boldsymbol{\gamma}}'\boldsymbol{B}_{k,l}\right)\left\{\frac{\widetilde{\boldsymbol{S}}_2\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)}{\widetilde{S}_0\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)} - \frac{\widetilde{\boldsymbol{S}}_1\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)\widetilde{\boldsymbol{S}}_1\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)'}{\widetilde{S}_0\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)^2}\right\}\mathrm{d}N_{k,l}(t)\right]^{-1}$$
$$\times\left[\int_t\left\{\boldsymbol{X}_{i,j} - \frac{\widetilde{\boldsymbol{S}}_1\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)}{\widetilde{S}_0\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)}\right\}\left\{\mathrm{d}N_{i,j}(t) - \frac{Y_{i,j}(t)\exp\left(\widetilde{\boldsymbol{\beta}}'\boldsymbol{X}_{i,j}\right)\sum_{l=1}^{J}\sum_{k=1}^{n^{(l)}}\xi_{k,l}V_{k,l}w_{k,l}^{(2)}\exp\left(\widetilde{\boldsymbol{\gamma}}'\boldsymbol{B}_{k,l}\right)\mathrm{d}N_{k,l}(t)}{\widetilde{S}_0\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)}\right\}\right].$$

$$IF^{(2)}_{i,j}\left\{\mathrm{d}\widetilde{\Lambda}_0(t)\right\} = \left\{\widetilde{S}_0\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)\right\}^{-1}\mathrm{d}N_{i,j}(t) + \left\{\sum_{l=1}^{J}\sum_{k=1}^{n^{(j)}}\xi_{k,l}V_{k,l}\exp\left(\widetilde{\boldsymbol{\gamma}}'\boldsymbol{B}_{k,l}\right)\widetilde{H}_{k,l}(t)\boldsymbol{B}'_{k,l}\right\}\boldsymbol{IF}^{(2)}_{i,j}\left(\widetilde{\boldsymbol{\gamma}}\right),$$

and $IF^{(3)}_{i,j}\left\{\mathrm{d}\widetilde{\Lambda}_0(t)\right\} = \widetilde{H}_{i,j}(t) + \left\{\sum_{l=1}^{J}\sum_{k=1}^{n^{(j)}}\xi_{k,l}V_{k,l}\exp\left(\widetilde{\boldsymbol{\gamma}}'\boldsymbol{B}_{k,l}\right)\widetilde{H}_{k,l}(t)\boldsymbol{B}'_{k,l}\right\}\boldsymbol{IF}^{(3)}_{i,j}\left(\widetilde{\boldsymbol{\gamma}}\right)$, with $\widetilde{H}_{i,j}(t) = -\widetilde{S}_0\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)^{-1}\mathrm{d}\widetilde{\Lambda}_0(t)\left\{\widetilde{\boldsymbol{S}}_1\left(t;\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\beta}}\right)'\widetilde{\boldsymbol{Z}}_{i,j} + \widetilde{K}_{i,j}(t)\right\}$, and $\widetilde{K}_{i,j}(t) = w^{(2)}_{i,j}Y_{i,j}(t)\exp\left(\widetilde{\boldsymbol{\beta}}'\boldsymbol{X}_{i,j}\right)$.

Finally, for any $s \in \{2,3\}$, $IF^{(s)}_{i,j}\left\{\int_{\tau_1}^{\tau_2}\mathrm{d}\widetilde{\Lambda}_0(t)\right\} = \int_{\tau_1}^{\tau_2}IF^{(2)}_{i,j}\left\{\mathrm{d}\widetilde{\Lambda}_0(t)\right\}$ and $IF^{(s)}_{i,j}$ $\left\{\widehat{\pi}^*\left(\tau_1,\tau_2;\boldsymbol{x}\right)\right\} = \left\{\frac{\partial\widetilde{\pi}(\tau_1,\tau_2;\boldsymbol{x})}{\partial\boldsymbol{\beta}}_{|\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}\right\}\boldsymbol{IF}^{(s)}_{i,j}\left(\widetilde{\boldsymbol{\beta}}\right) + \left[\frac{\partial\widetilde{\pi}(\tau_1,\tau_2;\boldsymbol{x})}{\partial\left\{\int_{\tau_1}^{\tau_2}\mathrm{d}\Lambda_0(t)\right\}}_{|\mathrm{d}\Lambda_0(t)=\mathrm{d}\widetilde{\Lambda}_0(t)}\right]IF^{(s)}_{i,j}\left\{\int_{\tau_1}^{\tau_2}\mathrm{d}\widetilde{\Lambda}_0(t)\right\}$.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10985-024-09621-2.

**Acknowledgements** We thank Dr. Barry Graubard and Dr. Edmund Jones for insightful comments and Dr. Arash Etemadi and GEMShare for documenting and sharing data from the Golestan Cohort. We are grateful to the Reviewers and Editors of Lifetime Data Analysis for valuable suggestions. This work was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster.

**Data availability** R code and functions used for the simulations in Section 7 are available in the Supplementary Information of this article and on GitHub at https://github.com/Etievant/CaseCohort. The R package CaseCohortCoxSurvival integrating the various functions on GitHub is available on CRAN at https://CRAN.R-project.org/package=CaseCohortCoxSurvival. We are not authorized to release the clinical data used in Sect. 8.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest. Dr Mitchell H. Gail is an Associate Editor of Lifetime Data Analysis.

## References

Andersen PK, Gill RD (1982) Cox's regression model for counting processes: a large sample study. Ann Stat 10(4):1100–1120

Barlow WE (1994) Robust variance estimation for the case-cohort design. Biometrics 50(4):1064–1072. https://doi.org/10.2307/2533444

Borgan Ø, Langholz B, Samuelsen SO, Goldstein L, Pogoda J (2000) Exposure stratified case-cohort designs. Lifetime Data Anal 6(1):39–58. https://doi.org/10.1023/a:1009661900674

Borgan Ø, Breslow NE, Chatterjee N, Gail MH, Scott A, Wild CJ (eds) (2017) Handbook of statistical methods for case-control studies. Chapman and Hall/CRC, London. https://doi.org/10.1201/9781315154084

Breslow NE (1974) Covariance analysis of censored survival data. Biometrics 30(1):89–99. https://doi.org/10.2307/2529620

Breslow NE, Lumley T (2013) Semiparametric models and two-phase samples: applications to Cox regression. From probability to statistics and back: high-dimensional models and processes—a Festschrift in honor of Jon A. Wellner 9:65–78. https://doi.org/10.1214/12-IMSCOLL906

Breslow NE, Wellner JA (2007) Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. Scand J Stat 34(1):86–102. https://doi.org/10.1111/j.1467-9469.2006.00523.x

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M (2009a) Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. Stat Biosci 1(1):32–49. https://doi.org/10.1007/s12561-009-9001-6

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M (2009b) Using the whole cohort in the analysis of case-cohort data. Am J Epidemiol 169(11):1398–1405. https://doi.org/10.1093/aje/kwp055

Chen K (2001) Generalized case-cohort sampling. J R Stat Soc Ser B (stat Methodol) 63(4):791–809

Deville JC (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. Surv Methodol 25(2):193–203

Ding J, Lu TS, Cai J, Zhou H (2017) Recent progresses in outcome-dependent sampling with failure time data. Lifetime Data Anal 23(1):57–82. https://doi.org/10.1007/s10985-015-9355-7

Graubard BI, Fears TR (2005) Standard errors for attributable risk for simple and complex sample designs. Biometrics 61(3):847–855. https://doi.org/10.1111/j.1541-0420.2005.00355.x

Gray RJ (2009) Weighted analyses for cohort sampling designs. Lifetime Data Anal 15(1):24–40. https://doi.org/10.1007/s10985-008-9095-z

Islami F, Kamangar F, Nasrollahzadeh D, Aghcheli K, Sotoudeh M, Abedi-Ardekani B, Merat S, Nasseri-Moghaddam S, Semnani S, Sepehr A, Wakefield J, Møller H, Abnet CC, Dawsey SM, Boffetta P, Malekzadeh R (2009) Socio-economic status and oesophageal cancer: results from a population-based case-control study in a high-risk area. Int J Epidemiol 38(4):978–988. https://doi.org/10.1093/ije/dyp195

Jones E (2018) cchs: an R package for stratified case-cohort studies. R J 10(1):484. https://doi.org/10.32614/RJ-2018-012

Jones E (2020) cchs: Cox model for case-cohort data with stratified subcohort-selection (0.4.2) [Computer software]. https://CRAN.R-project.org/package=cchs

Kalbfleisch JD, Prentice RL (2011) The statistical analysis of failure time data. Wiley, Hoboken

Keogh RH, Seaman SR, Bartlett JW, Wood AM (2018) Multiple imputation of missing data in nested case-control and case-cohort studies. Biometrics 74(4):1438–1449. https://doi.org/10.1111/biom.12910

Kim S, Zeng D, Cai J (2018) Analysis of multiple survival events in generalized case-cohort designs. Biometrics 74(4):1250–1260. https://doi.org/10.1111/biom.12923

Langholz B, Jiao J (2007) Computational methods for case-cohort studies. Comput Stat Data Anal 51(8):3737–3748. https://doi.org/10.1016/j.csda.2006.12.028

Lin DY (2000) On fitting Cox's proportional hazards models to survey data. Biometrika 87(1):37–47. https://doi.org/10.1093/biomet/87.1.37

Lumley T (2021) survey: analysis of complex survey samples (4.1–1) [computer software]. https://CRAN.R-project.org/package=survey

Lumley T, Shaw PA, Dai JY (2011) Connections between survey calibration estimators and semiparametric models for incomplete data. Int Stat Rev 79(2):200–220

Mark SD, Katki HA (2006) Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (nested) cohort studies with missing case data. J Am Stat Assoc 101(474):460–471. https://doi.org/10.1198/016214505000000952

Pfeiffer RM, Gail MH (2017) Absolute risk: methods and applications in clinical management and public health. Chapman and Hall/CRC, London. https://doi.org/10.1201/9781315117539

Pourshams A, Khademi H, Malekshah AF, Islami F, Nouraei M, Sadjadi AR, Jafari E, Rakhshani N, Salahi R, Semnani S, Kamangar F, Abnet CC, Ponder B, Day N, Dawsey SM, Boffetta P, Malekzadeh R (2010) Cohort profile: the Golestan cohort study—a prospective study of oesophageal cancer in northern Iran. Int J Epidemiol 39(1):52–59. https://doi.org/10.1093/ije/dyp161

Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73(1):1–11. https://doi.org/10.2307/2336266

Pugh M, Robins J, Lipsitz S, Harrington D (1993) Inference in the Cox proportional hazards model with missing covariate data

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89(427):846–866. https://doi.org/10.2307/2290910

Samuelsen SO, Ånestad H, Skrondal A (2007) Stratified case-cohort analysis of general cohort sampling designs. Scand J Stat 34(1):103–119. https://doi.org/10.1111/j.1467-9469.2006.00552.x

Sharp SJ, Poulaliou M, Thompson SG, White IR, Wood AM (2014) A review of published analyses of case-cohort studies and recommendations for future reporting. PLoS ONE 9(6):e101176. https://doi.org/10.1371/journal.pone.0101176

Shin YE, Pfeiffer RM, Graubard BI, Gail MH (2020) Weight calibration to improve the efficiency of pure risk estimates from case-control samples nested in a cohort. Biometrics 76(4):1087–1097. https://doi.org/10.1111/biom.13209

Therneau TM, Li H (1999) Computing the Cox model for case cohort designs. Lifetime Data Anal 5(2):99–112. https://doi.org/10.1023/a:1009691327335

Therneau TM, Lumley T, Elizabeth A, Cynthia C (2023) Survival: survival analysis (3.5–3) [Computer software]. https://CRAN.R-project.org/package=survival

Tsiatis AA (2006) Semiparametric theory and missing data. Springer, Berlin

Xu Y, Kim S, Zhang MJ, Couper D, Ahn KW (2022) Competing risks regression models with covariates-adjusted censoring weight under the generalized case-cohort design. Lifetime Data Anal 28(2):241–262. https://doi.org/10.1007/s10985-022-09546-8