Rebora *et al. BMC Medical Research Methodology* (2016) 16:5
DOI 10.1186/s12874-015-0103-1

BMC Medical Research Methodology

TECHNICAL ADVANCE Open Access

Crude incidence in two-phase designs in the presence of competing risks

Paola Rebora[1*], Laura Antolini[1], David V. Glidden[2] and Maria Grazia Valsecchi[1]

Abstract

Background: In many studies, some information might not be available for the whole cohort, some covariates, or even the outcome, might be ascertained in selected subsamples. These studies are part of a broad category termed two-phase studies. Common examples include the nested case-control and the case-cohort designs. For two-phase studies, appropriate weighted survival estimates have been derived; however, no estimator of cumulative incidence accounting for competing events has been proposed. This is relevant in the presence of multiple types of events, where estimation of event type specific quantities are needed for evaluating outcome.

Methods: We develop a non parametric estimator of the cumulative incidence function of events accounting for possible competing events. It handles a general sampling design by weights derived from the sampling probabilities. The variance is derived from the influence function of the subdistribution hazard.

Results: The proposed method shows good performance in simulations. It is applied to estimate the crude incidence of relapse in childhood acute lymphoblastic leukemia in groups defined by a genotype not available for everyone in a cohort of nearly 2000 patients, where death due to toxicity acted as a competing event. In a second example the aim was to estimate engagement in care of a cohort of HIV patients in resource limited setting, where for some patients the outcome itself was missing due to lost to follow-up. A sampling based approach was used to identify outcome in a subsample of lost patients and to obtain a valid estimate of connection to care.

Conclusions: A valid estimator for cumulative incidence of events accounting for competing risks under a general sampling design from an infinite target population is derived.

Keywords: Two-phase design, Competing risks, Crude incidence, Case-control, Case-cohort, Missing data, Subdistribution hazard

Background

In many longitudinal studies, some information might not be measured/available for the whole cohort, in fact biomarkers/additional covariates, or even outcome, might be ascertained only in selected subsamples. These studies are part of a broad category termed two-phase studies [1], in fact they imply two sampling phases: the first one being usually a random sample from the target population, ending up in the entire cohort (phase I sample), and the second one applying some kind of sampling (e.g. efficient or of convenience) to collect additional information or the selection of subjects with no missing data (phase II sample). Common examples of efficient second phase sampling include the nested case-control and the case-cohort designs [2–5]. In other situations the outcome itself is collected only for a subsample [6]. Two-phase sampling, or more generally, multiphase sampling, is a general design that includes any valid probability sample of the data, in which each subsampling can depend on all the currently observed data at each step [7]. The actual sampling probabilities will depend on the specific design. The acknowledgement of these sampling phases, even in the commonly applied designs, can be very useful to improve efficiency and to allow flexibility in the analysis (e.g. different time-scales or different models can be applied) by using information available for the whole cohort [8].

Efficient designs are particularly useful to identify new biomarkers when the combination between large cohorts

*Correspondence: paola.rebora@unimib.it
[1]Center of Biostatistics for Clinical Epidemiology, School of Medicine and Surgery, University of Milano-Bicocca, via Cadore 48, 20900 Monza, Italy
Full list of author information is available at the end of the article

BioMed Central

© 2016 Rebora et al. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

and expensive new technologies make it infeasible to measure the biomarkers on the entire cohort. The Women's Health Initiative program, for example, stored serum and plasma from participants and used them for specialized studies [9]. Also the Cardiovascular Health Study collected DNA from most participants to study different genetic factors underlying cardiovascular or other diseases and only subsets of the cohort have been genotyped in different projects [10].

In our first motivating clinical example, the aim was to evaluate the role of different genetic polymorphisms on treatment failure due to relapse in childhood acute lymphoblastic leukemia (ALL) using clinical information and biological samples available from a clinical trial that enrolled nearly 2000 patients. In this situation, a parsimonious use of these specimens motivated the choice of an efficient/optimal two-phase sampling design [11]. We present also a further application where the aim was to estimate engagement to care of HIV patients in resource limited settings. Here the outcome itself was missing in a group of patients due to possibly informative loss to follow-up. The outcome was tracked in a random sample of those lost to follow-up to obtain a valid estimate of engagement to care [12].

For two-phase studies, appropriate weighted survival estimates have been derived, both in the presence of additional covariates measured in the second phase [7, 13, 14], as well as in cohorts where the outcome/follow-up is not available for everyone [15]. A Cox model adapted for two-phase designs has also been derived [14]. However no estimator of cumulative incidence accounting for competing events in the general framework of two-phase designs has been proposed, while it has been developed for specific designs, such as nested case-control studies [16, 17]. This is relevant in the presence of multiple types of events, such as relapse and (toxic) death in cancer patients, as in the motivating examples presented here, where estimation of event type specific quantities are needed for evaluating outcome.

The aim of this paper is to develop a non parametric estimator of the crude incidence of events accounting for possible competing events in the general framework of two-phase designs, where subgroups of analysis might be defined according to explanatory variables ascertained in the phase II sample, or the outcome itself assessed only in the second phase sample.

In the Methods section we propose a weighted crude incidence estimator for application in two-phase designs. The theoretical properties of the proposed method are derived in appendix and investigated through simulations under different scenarios, which results are reported in Results section. In this section we also report the examples on childhood ALL and HIV patients. Conclusions is dedicated to the discussion.

Methods

Notation and basics

Let T be the failure time variable and suppose there are K possible causes of failure denoted by $\varepsilon = 1, 2, \ldots K$. Let the cause-specific hazard function of the k^{th} event be:

$$\lambda_k(t) = \lim_{dt \to 0} \frac{1}{dt} P(t \leq T < t + dt; \varepsilon = k | T \geq t)$$

and $\Lambda_k(t) = \int_0^t \lambda_k(s) ds$. Define

$$F_k(t) = P(T \leq t; \varepsilon = k) \tag{1}$$

as the probability that a failure due to cause k occurs by time t, that is the quantity that we aim to estimate. Define also $S(t) = P(T > t) = 1 - \sum_k F_k(t)$ as the probability of surviving from any cause of failure.

A convenient representation of the crude incidence function (1) as product limit estimator naturally arises starting from the subdistribution hazard introduced by Gray [18] and defined as:

$$\begin{aligned} \lambda_k^*(t) = \lim_{dt \to 0} \frac{1}{dt} P\{t \leq T < t + dt; \varepsilon \\ = k | T \geq t \cup (T < t; \varepsilon \neq k)\} \end{aligned} \tag{2}$$

This hazard has been shown to be very useful to compare the crude cumulative hazard functions in different groups, since it restores a one-to-one relationship between the hazard and the cumulative probability of a particular failure type: $F_k(t) = 1 - \exp\{-\Lambda_k^*(t)\} = 1 - \prod_{s \leq t} [1 - \Lambda_k^*(ds)]$, with $\Lambda_k^*(t) = \int_0^t \lambda_k^*(s) ds$ and where the product integral notation $\prod$ is used to suggest a limit of finite products $\prod$ [18, 19]. Of note, the one-to-one relationship between the hazard and the cumulative probability is not satisfied from the cause-specific hazard in the presence of competing events [20]. The subdistribution hazard can be thought as the hazard of an artificial variable $T_k^* = T \cdot I\{\varepsilon = k\} + \infty \cdot I\{\varepsilon \neq k\}$ that extends to infinity the time to event k when another competing event is observed. In fact, for any finite t, $T_k^* \leq t$ is equivalent to $T \leq t$ and $\varepsilon = k$; thus, given definition (1), $P(T_k^* \leq t) = F_k(t)$. The definition of T_k^* is consistent with the argument that when an event other than k occurs as first, the latter will never be observed as first and thus the corresponding time is infinity.

Let $(T_i, \varepsilon_i, C_i, Z_i)$, with $i = 1 \ldots N$, be N independent replicates of (T, ε, C, Z), where C is the censoring time and Z a vector of covariates. We will refer to these N subjects as the phase I sample. Define $X = \min(T, C)$ and $\Delta = I(T \leq C)$. We will assume that failure and censoring times are conditionally independent, $T \perp C | Z$. Let $Y_i(t) = I(X_i \geq t)$, $N_{ik}(t) = I(X_i \leq t, \Delta_i \varepsilon_i = k)$ and $N_{i\cdot}(t) = \sum_{k=1}^K N_{ik}(t)$, where $I(\cdot)$ is the indicator function. Define $G(t) = P(C > t)$ as the probability to remaining uncensored up to t.

Suppose that complete information on $(X_i, \Delta_i\varepsilon_i, Z_i)$ is available only for a subset $n < N$ of subjects drawn based on a possibly complex sampling design and let ξ_i indicate whether subject i is selected into this sample. We will refer to the $n = \sum_i \xi_i$ subjects as the phase II sample, even if multiple phases of sampling could actually be involved to obtain the final complete sample [7]. Let $\pi_i = P(\xi_i = 1|X_i, \Delta_i\varepsilon_i, Z_i)$ being the inclusion probability of subject i for the phase II sample, conditional on being selected at the first phase. In a random sample this probability is equal for every subject. However sampling is often stratified on some variables to increase efficiency; in this case, the probability to be selected for the phase II sample is common for all subjects in the same stratum and differs between strata. In particular, it is usually higher for the more informative strata (e.g. strata including subjects with the event of interest as in case-control studies). For nested case-control designs the sampling probability of cases will be 1, while the one of controls might be derived as the probability that individual i is ever selected as control, following Samuelsen [4]. We denote the pairwise sampling probability for any two subjects (i, j, with $i \neq j$) by $\pi_{ij} = P(\xi_i = 1, \xi_j = 1|X_i, \Delta_i\varepsilon_i, Z_i, X_j, \Delta_j\varepsilon_j, Z_j)$. As commonly assumed in survey theory, the sampling method should have the following properties: the sampling probabilities π_i and π_{ij} must be non zero for all i, j in the population and must be known for each i, j in the sample [7].

Incidence estimation in the presence of competing risks

Overall survival/incidence estimate

Under a two-phase design it is common to be interested in estimating survival in subgroups related to variables ascertained only in phase II sample (i.e. biomarkers). Another possible situation is that, instead of covariates, the outcome itself is not available for the whole cohort. Thus, in both cases an estimate of the incidence of event using only the phase II sample is very useful. The total number of events of type k up to t and the total number of persons at risk at time t for the entire phase I sample can be estimated from the phase II sample (accounting for the sample design) by $\hat{N}_{\cdot k}(t) = \sum_{i=1}^{N}[\xi_i N_{ik}(t)/\pi_i]$ and $\hat{Y}_{\cdot}(t) = \sum_{i=1}^{N}[\xi_i Y_i(t)/\pi_i]$, respectively. Note that these estimates are valid under general sampling designs, where π_i and π_{ij}, the so-called 'design weights', are known for the observations actually sampled [21].

The estimate of the overall survival has been shown by several authors in different contexts of complex sampling [13–15]:

$$\hat{S}(t) = \prod_{s\leq t}\left[1 - \hat{\Lambda}(ds)\right] \tag{3}$$

where the overall hazard can be obtained by $\hat{\Lambda}(t) = \sum_{k=1}^{K}\hat{\Lambda}_k(t)$ and $\hat{\Lambda}_k(t) = \int_0^t \hat{N}_{\cdot k}(ds)/\hat{Y}_{\cdot}(s)$ [22]. It has been shown that $\sqrt{N}[\hat{\Lambda}(t) - \Lambda(t)]$ converges weakly to a zero-mean Gaussian process [14, 15].

Competing risk

The goal is to estimate the crude incidence of a given cause k, $F_k(t) = 1 - \prod_{s\leq t}\left[1 - \Lambda_k^*(ds)\right]$, using the phase II sample, which is also called subdistribution function and is the probability that a failure due to cause k occurs within t [23, 24]. The estimate of $\Lambda_k^*(t)$ is based on the count of events due to cause k and the count of subjects at risk for T_k^*, denoted by $\hat{Y}_{\cdot k}^*(s)$ (see Appendix A.1):

$$\hat{Y}_{\cdot k}^*(s) = \sum_{i=1}^{N}\frac{\xi_i}{\pi_i}Y_i(s) + \sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\left[\sum_{l\neq k}N_{il}(s^-)\cdot\hat{G}(s^-|X_i^-)\right] \tag{4}$$

The estimate of the cumulative subdistribution hazard in (2) can now be estimated, using only the phase II sample, by:

$$\hat{\Lambda}_k^*(t) = \int_0^t \frac{\hat{N}_{\cdot k}(ds)}{\hat{Y}_{\cdot k}^*(s)}. \tag{5}$$

Note the complement of $F_k(t)$ can be thought as the survival probability of T_k^* [18, 20, 25], thus a product limit type estimator can be directly derived as:

$$\hat{F}_k(t) = 1 - \prod_{s\leq t}\left[1 - \hat{\Lambda}_k^*(ds)\right] = 1 - \prod_{s\leq t}\left[1 - \frac{\hat{N}_{\cdot k}(ds)}{\hat{Y}_{\cdot k}^*(s)}\right] \tag{6}$$

Interestingly, this estimator is algebraically equivalent to the Aalen-Johansen type estimator, shown by [18] for random sampling, and in the Appendix A.2 for general sampling:

$$\hat{F}_k(t) = 1 - \prod_{s\leq t}\left[1 - \hat{\Lambda}_k^*(ds)\right] = \int_0^t \hat{S}(s^-)\hat{\Lambda}_k(ds) \tag{7}$$

It is easy to see that in the absence of competing events, $\hat{Y}_{\cdot k}^*(s)$ in (4) degenerates to the usual risk set $\hat{Y}_{\cdot}(s)$, thus $\hat{\Lambda}_k^*(t) = \hat{\Lambda}(t)$ and $\hat{F}_k(t)$ equals the complement of 1 of the weighted Kaplan-Meier estimator for two-phase studies [13]. Under no censoring, the weight $\hat{G}(s^-|X_i)$ becomes 1 and the risk set $Y_{\cdot k}^*$ is eroded in time only by events of type k, therefore $\hat{F}_k(t)$ degenerates into the proportion of events of type k estimated by the phase II sample (weighted number of events of type k out of the estimated total size of the cohort, phase I). If every subject in phase I is sampled ($\xi_i = 1\,\forall i$), then (5) becomes the standard subdistribution cumulative hazard [19, 20] and (6) the standard estimator of the crude incidence.

For simplicity of notation, in (6) we estimated the overall incidence regardless of covariates, but the estimator can

also be applied on subgroups defined by Z. The censoring probability $G(t)$ should also be estimated in subgroups defined by Z. The overall estimator is reasonable when we make the more restrictive assumption $T \perp C$, otherwise separate estimators conditional on Z would be more appropriate (and eventually an average, weighted on the frequencies of Z, between the conditional estimates).

Variance and confidence intervals

Following Breslow and Wellner [26], we can express $\sqrt{N}\left[\hat{\Lambda}^*_k(t) - \Lambda^*_k(t)\right] = \sqrt{N}\left[\tilde{\Lambda}^*_k(t) - \Lambda^*_k(t)\right] + \sqrt{N}\left[\hat{\Lambda}^*_k(t) - \tilde{\Lambda}^*_k(t)\right]$ where $\tilde{\Lambda}^*_k(t)$ represents the crude cumulative incidence estimator that we would have obtained if complete information $(X_i, \Delta_i\varepsilon_i, Z_i)$ was known for all the subjects in phase I sample $(i = 1 \dots N)$ [18]. The two terms are asymptotically independent [14, 26]. The first term converges weakly to a zero-mean Gaussian process [19] with covariance that we denote as $\sigma^2_{kI}(t)$. By the arguments in Appendix A.3, the second term converges weakly to a zero-mean Gaussian process with covariance $\sigma^2_{kII}(t)$. Hence, $\sqrt{N}\left[\hat{\Lambda}^*_k(t) - \Lambda^*_k(t)\right]$ converges weakly to a zero-mean Gaussian process with covariance being the sum of the contribution of each sampling phase: $\sigma^2_k(t) = \sigma^2_{kI}(t) + \sigma^2_{kII}(t)$. The first one represents the irreducible minimum uncertainty that would remain if everyone in phase I would be sampled and the second one accounts for the fact that complete information is available only in the phase II sample [13, 14, 26].

Each contribution to the variance can be estimated by the influence function approach [27]. The influence function of an estimator describes how the estimator changes when single observations are added or removed from the data and has the property that the difference between the estimate and the population quantity can be expressed as the sum of influence functions over all the subjects in the sample. By denoting with $z^*_{ik}(t)$ the influence function of subject i on $\hat{\Lambda}^*_k(t)$ we have that [28]:

$$\hat{\Lambda}^*_k(t) - \Lambda^*_k(t) = \sum_{i=1}^{N} z^*_{ik}(t) + o(1/\sqrt{N}) \tag{8}$$

The influence function of subject i on $\hat{\Lambda}^*_k(t)$ has been derived in Appendix A.4.

By using the Horvitz-Thomposon variance [29] on the weighted influence function, the contribution of the variance of phase II will be:

$$\begin{aligned}\hat{\sigma}^2_{kII}(t) &= \hat{var}\left[\sum_{i=1}^{N} \frac{\xi_i}{\pi_i} z^*_{ik}(t)\right] = \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{z^*_{ik}(t)\cdot z^*_{jk}(t)}{\pi_i \cdot \pi_j} - \frac{z^*_{ik}(t)\cdot z^*_{jk}(t)}{\pi_{ij}}\right]\end{aligned} \tag{9}$$

For phase I, the variance $\hat{\sigma}^2_{kI}(t)$ can also be estimated using (9) by setting sampling probabilities to 1 [13].

Given the one-to-one relationship between $F_k(t)$ and $\Lambda^*_k(t)$, the variance of the crude cumulative incidence (6) can now be estimated as:

$$\hat{var}\left[\hat{F}_k(t)\right] = \left[1 - \hat{F}_k(t)\right]^2 \cdot \hat{\sigma}^2_k(t) \tag{10}$$

In analogy with the survival estimate for two-phase designs, we derived confidence intervals for (6) on the logarithm scale by:

$$\exp\left\{\log[\hat{F}_k(t)] \pm q_{\alpha/2}\frac{1-\hat{F}_k(t)}{\hat{F}_k(t)}\hat{\sigma}_k(t)\right\} \tag{11}$$

where $q_{\alpha/2}$ denotes the $\alpha/2$ quantile of the standard Gaussian distribution.

Software

By using suitable weights for both study design and censoring, any software allowing for time dependent weights can be used to derive the modified risk set and to estimate the crude cumulative incidence function (6). These weights have been implemented in R in function `crprep` in the `mstate` package [30] and in STATA in the `stcrprep` function. However, any software can be used to derive the ingredients for the modified risk set in (4) and these can be used to estimate (6) and its variance by the Horvitz-Thompson approach.

The complete code to compute this estimate has been developed in R software [31] using the `survey` package [32] and is available at [33]. An example of the application of this function is given in the subsection Genotype ascertained on a subset of a clinical trial cohort.

Results

Simulations

Simulations protocol

We considered two competing events with independent latent times T_1 and T_2 and constant marginal hazard of 0.1, the crude incidences are then $F_1(t) = F_2(t) = \frac{1}{2}(1 - e^{-0.1t})$. We focused on the crude incidence of event 1 up to $t = 2$ units of time (i.e. years). This implies a fraction of 82 % with no events at $t = 2$ (administrative censoring) and a crude incidence of about 9 %. The independence between the latent times T_1 and T_2 is not restrictive given the non identifiability issue [34]. The censoring time followed an uniform distribution on ranges (0.5,30.5) and (0.5,10.5), leading to around 5 % and 15 % censored before $t = 2$, respectively.

We drew $B = 1000$ random first-phase samples of size $N = 1000$, from which we sampled a phase II sample according to different study designs:

1. random sample, with $n = 50, 100$ units;
2. case-control sampling: we randomly sampled $n/2$ individuals among those who experienced event 1

(cases) up to time 2 and $n/2$ individuals among the others (controls), with $n = 50, 100$ units;

3. stratified sampling: we considered the phase I sample divided into 4 strata defined by the variable $Z = \{0, 1\}$ (with frequencies 70 % and 30 % for $Z = 0$ and $Z = 1$, respectively) and the occurrence or not of event 1 up to time 2. The hazard rates were assumed to be 0.08 and 0.2 with $Z = 0$ and with $Z = 1$, respectively. An equal number of subjects ($n/4$) were sampled for each strata (balanced sampling), with $n = 50, 100$ units.
4. nested case-control design: we selected all cases and m controls for each case with no events at the time of event of the case, fixing $m = 1, 2$. Under this design we cannot fix a total sample size a priori, but we expect around 90 events and $90 \cdot m$ controls. Sampling probabilities for each included subject were derived according to Samuelsen [4].

B = 1000 was chosen in order to get a ± 5 % level of accuracy in the estimate of the crude incidence ($F_1(t), t > 0.3$) in about 95 % of the samples. For each sample, $\hat{F}_1(t)$ has been computed by (6), with $\hat{\Lambda}_1^*(dt)$ estimated by (5), and it has been compared with $F_1(t)$ in order to assess bias in each sample: $\hat{F}_{1b}(t) - F_1(t)$, $b = 1 \ldots B$. Bias has been computed and reported for 20 different time points $t = 0.1, 0.2, \ldots, 2$. For each simulation, we also computed standard error of $\hat{F}_1(t)$ according to (10) and the 95 % confidence interval (CI) of $\hat{F}_1(t)$ on the logarithm scale (11) to evaluate coverage and length.

Simulations results

Figure 1 compares the average of the estimated standard error of $\hat{F}_1(t)$ in each simulation with the empirical standard error at the 20 different times of observation in the four different scenarios with random censoring of 15 %. They were found to be very close in all scenarios, as expected.

Figure 2 reports the distribution of bias in each one of the 1000 simulated samples under random (panel a), case-control (panel b), stratified (panel c) and nested case-control sampling ($m = 1$, panel d). Bias fluctuates around 0 in each scenario, but it has more variability in the random sampling compared to other scenarios, resulting also in a higher mean value of absolute bias over the B simulations (still always lower than 0.2 %). The lower performance of the estimator in the first scenario is due to the fact that random sampling is not a convenient design in the simulated setting. In fact, in phase I cohort we expect around 90 events of type 1 (incidence of 9 % and sample size of 1000), thus if we randomly sample 100 subjects from the phase I cohort, we expect to observe only 9 events (in phase II sample). With such a small number of events, unbiasedness is in fact not sufficient to ensure a reasonable behaviour and to get enough information on event incidence. To address this issue, the other study designs (scenarios 2, 3, and 4) are indeed thought to guarantee to sample more events of type 1. Thus we recommend adopting efficient designs accounting for the event of interest. Relative and standardized biases were always lower than 6 % (data not shown). The mean square error, not shown, slightly increases with time, in fact variability is increasing in time (as confirmed by the empirical standard error of $\hat{F}_1(t)$ and by Fig. 2). The average length of the confidence interval was consistently increasing with time, ranging between 7 % and 12 % in the random sampling and between 2 % and 4 % in the case-control, stratified and nested case-control sampling (data not shown). This comparison underscores the advantages of a careful selection of the subsample.

Figure 3 reports results on coverage for the random, case-control, stratified and nested case-control sampling. The coverage was very close to the nominal value of 95 %, ranging mostly within a minimum of 94 % and a maximum of 97 %, except for very early times in the random setting.

In the same setting, we also considered a longer follow-up time, t = 50, with around 500 events of type 1 expected in phase I sample (under no censoring), and confirmed the performance of our estimator in a scenario with higher variability, with similar results for the different sampling schemes (data not shown).

Motivating examples

Genotype ascertained on a subset of a clinical trial cohort

A study on childhood ALL evaluated the role of a genetic polymorphism (glutathione S-transferase-θ, GST-T1) on treatment failure due to relapse (in different sites), in the presence of a competing event (toxic death). GST-T1 is a common genetic polymorphism in Caucasians, with 13–26 % of individuals displaying a homozygous deletion of the gene (null genotype). Subjects carrying the null variants fail to express the GST-T1 enzyme, that is involved in drug metabolism. Clinical information were available for a cohort of 1999 consecutive patients (mainly European Caucasians, aged between 1 and 17 years, median age: 5 years) newly diagnosed with ALL in the Italian Associazione Italiana di Ematologia Pediatrica centers between September 2000 and July 2006. Biological samples stored at diagnosis were available, but genotype was ascertained only in a subgroup (phase II sample) for an efficient use of specimens [11, 13]. The interest was to evaluate incidence at different relapse sites by GST-T1, that can only be estimated using phase II data. In order to select the subgroup to be genotyped we adopted an optimal strategy that is carefully described in [11, 13]. Briefly, sampling was done after classifying patients into 6 strata according to the event of interest (relapse/no relapse) and to 3 groups, defined by prognostic features in the treatment protocol, that modulate the intensity of treatment, we will call them

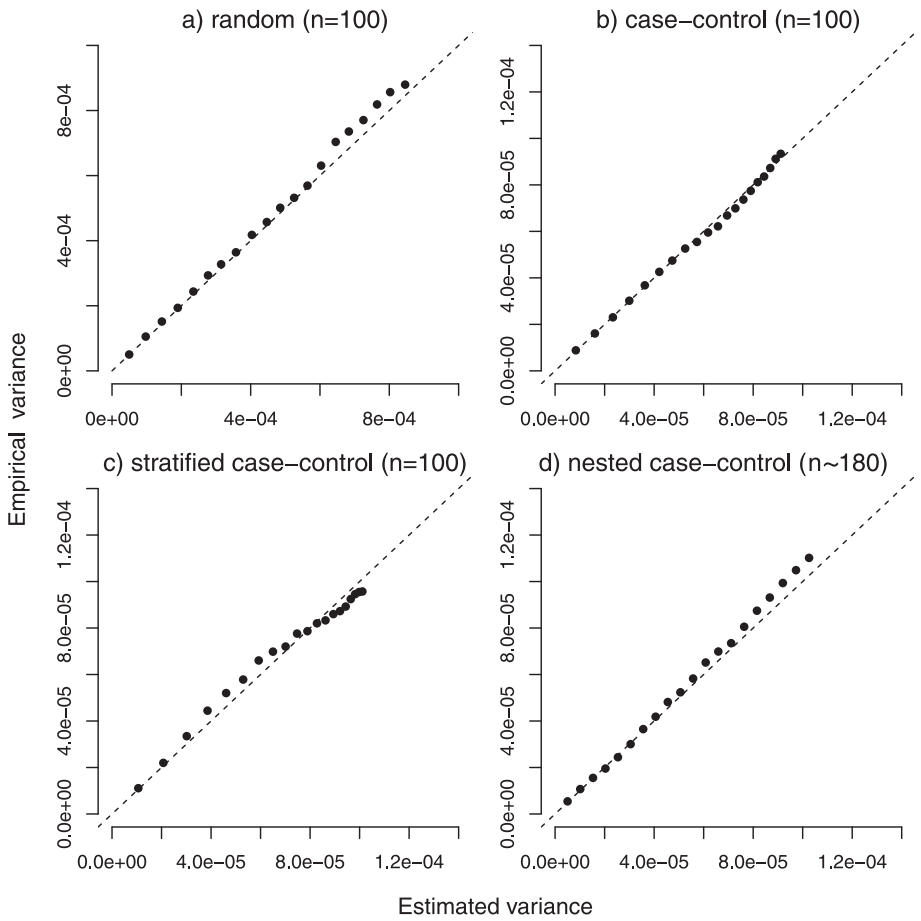

Fig. 1 Estimated and empirical variances. Comparison between estimated and empirical variance under random (panel **a**), case-control (panel **b**), stratified (panel **c**) and nested case-control sampling ($m = 1$, panel **d**). For the first three scenarios a sample size of 100 was used, while in the nested case-control sampling a mean of 180 subjects was considered. Data were subject to a random censoring of 15 % (plus administrative censoring). Dashed lines represent the main bisector corresponding to the equality between estimated and empirical reference

treatment protocols (Table 1). Strata were not defined based on the competing event death due to toxicity- 58 events - for efficiency reasons given that the event of interest was relapse. Patients were sampled at random without replacement from the 6 strata, with the sampling from each stratum conducted independently (stratified sampling) and with higher probability in the more informative strata according to an optimal design [13]. The full cohort of 1999 patients represents the phase I sample, for which clinical information are available, while genotype is ascertained in the phase II sample only ($n = 601$).

Relapses were classified according to the site, in particular we distinguished relapses involving bone-marrow (BM) from the others (extramedullary). We estimated the crude incidence of BM relapse by GST-T1 deletion using (6) and (10) and found higher relapse incidence for patients with GST-T1 deletion, with 5-year crude incidence of 19.3 % (95 % CI: 13.4 − 27.7 %) versus 12.4 % (95 % CI: 10.7 − 14.4 % for non deleted patients, Fig. 4 panel a). This was derived accounting for the competing risk of other sites of relapse as well as for death due to toxicity.

We report here the R code used to compute these estimates:

```
library(survey)
d.std<-twophase(id=list(~upn,~upn),
subset=~!is.na(GST_T), strata=list
(NULL,~interaction(rel,elfin)),data=dat)
GSTse<-svycr(Surv(time,event>0)~GST_T,
etype="BMrelapse",d.std,se=TRUE)
```

The `twophase` function in the `survey` package describes the design and produces a survey object [32]. The `svycr` function, available at [33], performs the estimate of crude incidence by the influence approach and uses 3 variables: `time` is the time of event, `event` the censoring indicator (1 if an event of any type is observed and 0 otherwise) and `BMrelapse` indicates whether a BM relapse is observed or not. Details on the `survey` package can be found in [7, 32].

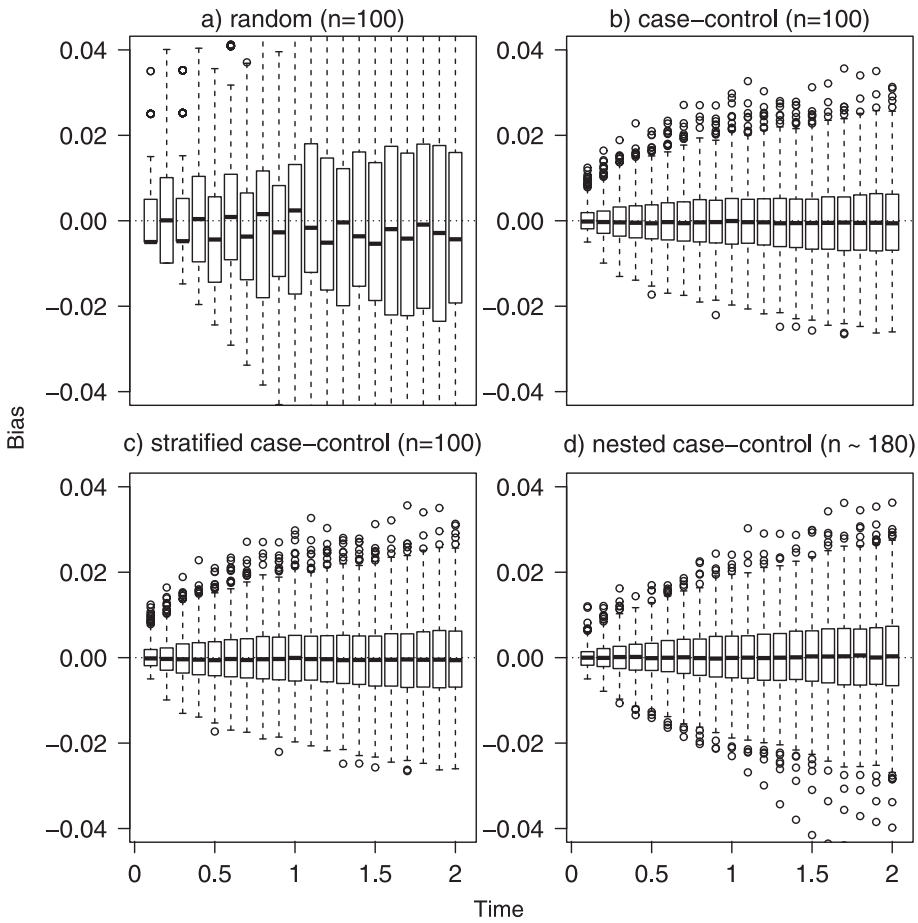

Fig. 2 Bias distribution. Distribution of bias in the 1000 simulated samples under random (panel **a**), case-control (panel **b**), stratified (panel **c**) and nested case control sampling ($m = 1$, panel **d**). For the first three scenarios a sample size of 100 was used, while in the nested case-control sampling a mean of 180 subjects was considered. Data were subject to a random censoring of 15 % (over administrative censoring). The box represent the first and third quartile, the black line the median and the empty dots represent outliers defined as bias more than 1.5 times the interquartile range above the third quartile (or more than 1.5 times the interquartile range below the first quartile). Dotted lines represent the reference for bias (bias equal to 0)

The right panel of Fig. 4 represents the incidence of extramedullary relapses by GST-T1 showing that the difference in relapse incidence between GST-T1 deleted and other patients is mainly due to relapse involving the BM, that represents the most relevant type of relapse in childhood ALL. A Cox model adapted for two-phase design [14], when applied to the cause specific hazard of BM relapse, gives an hazard ratio (HR) of 1.53 (95 % CI 0.98–2.37) for GST-T1 deleted patients versus non deleted; after adjusting for relevant factors (treatment protocol, gender, age), the HR dropped to 1.38 (95 % CI 0.90–2.13). For extramedullary relapses the HR was 1.22 (95 % CI 0.60–2.49). Of note, in order to compare patients with and without deletion of the GST-T1 gene, we used a cause-specific model, thus we actually compared the cause-specific hazard of relapse. In fact, a subdistribution model accounting for the two-phase design is not available. This would be useful to compare the actual incidences of relapse in the two groups, however the cause-specific model is still very useful to address the impact of the genotype on relapse by an aetiological point of view.

Outcome ascertained on a subset of patients lost to follow-up

In the evaluation of the effectiveness of the global effort to provide antiretroviral therapy (ART) for HIV-infected patients in resource limited settings, the estimate of the number of patients who continue to access care after starting ART is essential. This estimate is hampered, however, by the fact that some patients die shortly after their last visit to clinic - a group of individuals who cannot be considered as "stopping care" nor censored for the event of stopping care. In addition, the number of patients who are starting care is large and a high fraction have unknown outcomes (i.e., are lost to follow-up), generating informative censoring. Given that lost patients could reasonably be not in care, but they could also have changed clinic or

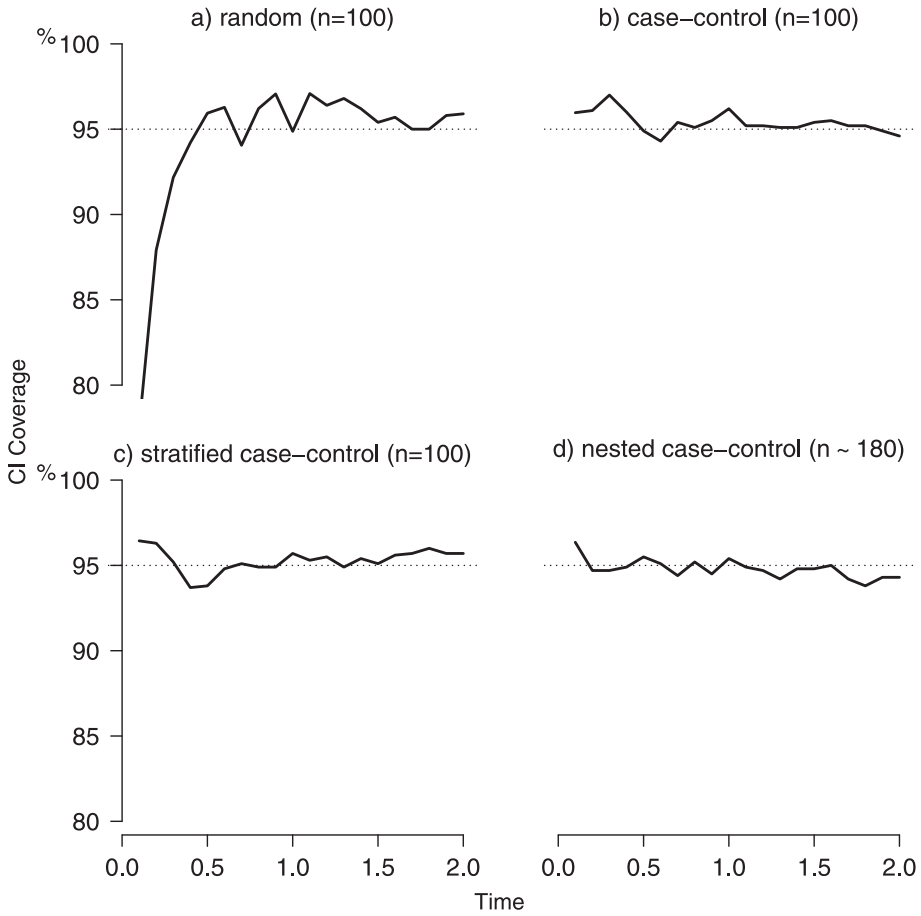

Fig. 3 Coverage of confidence intervals. Simulation results for the coverage of confidence intervals (CI) under random (panel **a**), case-control (panel **b**), stratified (panel **c**) and nested case-control sampling ($m = 1$, panel **d**). For the first three scenarios a sample size of 100 was used, while in the nested case control sampling a mean of 180 subjects was considered. Data were subject to a random censoring of 15 % (over administrative censoring). Dotted lines represent the reference for CI coverage (nominal 95 %)

be dead, one approach to obtain outcomes estimates has been to identify a numerically small, but random, sample of those who are lost [15], intensively seeking their outcomes, and using them to correct outcomes among the lost.

Table 1 Distribution of phase I (N_s) and II (n_s) samples in the 6 strata and sampling fractions expressed as percentages in parenthesis for Phase II

	Treatment protocol			
	Standard	Medium	High	
		n_s/N_s(%)		Total
No relapse	54/487 (11.1)	193/987 (19.6)	109/219 (49.8)	356/1693
Relapse	21/28 (75.0)	147/186 (79.0)	77/92 (83.7)	245/306
Total	75/515	340/1173	186/311	601/1999

To illustrate, a cohort of 13,321 HIV-infected adult patients, who initiated ART treatment, were followed from ART initiation to either death, disengagement or administrative database closure (see Fig. 5). Among them, 2451 patients were lost to follow-up [35], defined as not being seen at the clinic for at least 90 days (after the last return visit). A tracker went into the community to determine the outcome of a random subsample of 428 among the 2451 lost patients and got information on 306 patients (110 patient were found to be in care in other clinics, 80 died while in care and 116 were found to be not treated/disengaged) [12]. The 10,870 patients no lost to follow-up and the 306 tracked patients can been considered as the second phase sample of the whole cohort, stratified on lost to follow-up.

We used the methods developed in the Methods section to estimate crude cumulative incidence, where the 306 tracked patients represented the 2451 lost patients by the

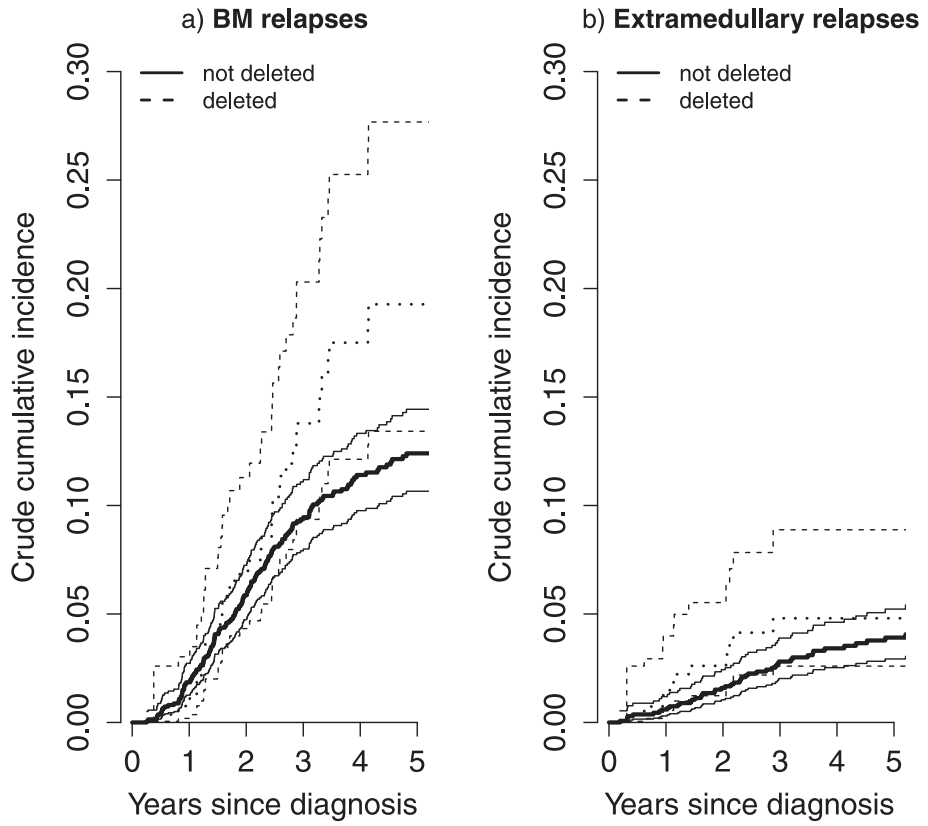

Fig. 4 Crude cumulative incidence of relapse in ALL data. Panel **a** reports the crude cumulative incidence estimate of relapse involving the bone marrow in patients with normal (estimate and confidence limits reported with solid lines) and deleted (dashed lines) GST-T1 gene. Panel **b** reports the crude cumulative incidence estimate of relapse in extramedullary sites in patients with normal (solid lines) and deleted (dashed lines) GST-T1 gene

sampling probability 306/2451, while the other 10,870 had sampling weight one. The crude incidence estimate of disengagement is reported in Fig. 6, the curve starts to rise after 90 days from ART start, that is the earliest possible time of disengagement, by definition. At 1 year, disengagement resulted 6.8 % (CI 95 % 5.7–8.2 %). This was subject to a strong influence of the competing event death in care that resulted 7.7 % (CI 95 % 6.8–8.9 %) at 1 year. A naïve (but less expensive) approach to deal with informative censoring would be to treat all lost patients as events or, contrarily, as censored observations. We plotted the two corresponding curves in Fig. 6, obtaining estimates of crude incidence at 1 year since ART treatment of 18.5 % and 0.2 %, respectively. We can consider that the true incidence will lie between these two estimates (that are however quite far in this context), as in fact it does

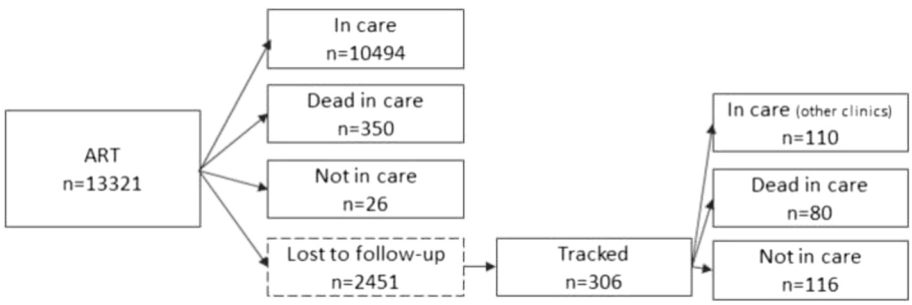

Fig. 5 Outcome in the cohort of HIV patients. Outcome of the 13,321 HIV patients who initiated ART treatment

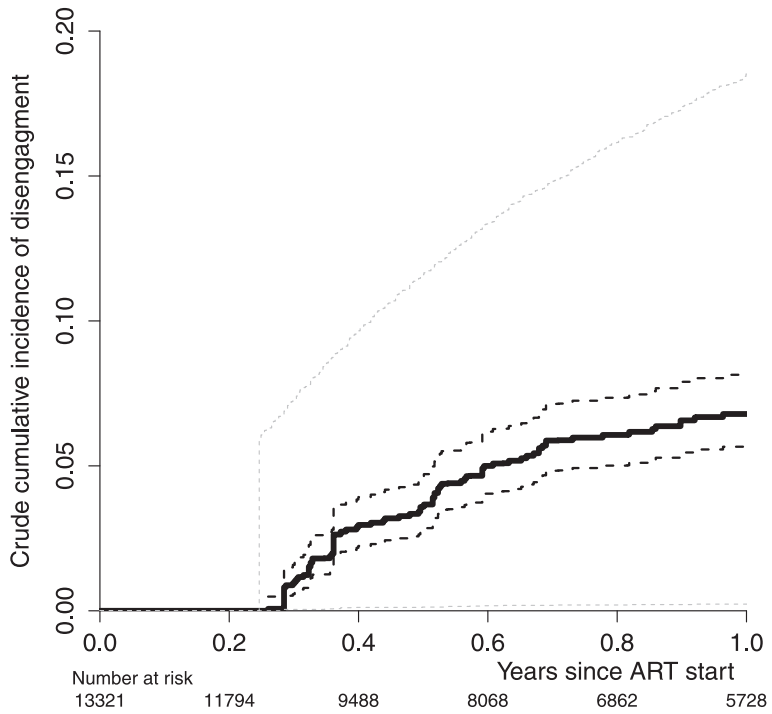

Fig. 6 Crude cumulative incidence of disengagement in the HIV data. Crude cumulative incidence of disengagement of the 13,321 HIV patients who initiated ART treatment (black line) with confidence intervals (dashed lines). In the bottom part of the plot the number of patients at risk in time is reported (weighed to represent the whole cohort of 13,321). The dashed grey lines report the crude cumulative incidence of disengagement computed by treating all lost patients as event or censored observation, respectively

the estimate we got by tracking a random sample of lost patients and using the proposed estimator.

Availability of supporting data

The R code to compute the proposed estimate of crude cumulative incidence is available at [33]. The results of simulations presented in the Simulations section and the related code are also available at [33].

Conclusions

We have derived an estimator for cumulative incidence of events based on the subdistribution hazard accounting for competing risks under a general sampling design from an infinite target population. The estimator shows good performance in simulations under different scenarios and the variance, derived by the influence function of the subdistribution hazard and the Horvitz-Thompson theory, was very close to the empirical variance, therefore we expect it to be very close to the one obtainable by replicate weights (e.g. bootstrap) [7]. Confidence intervals, derived on the log scale, provided good coverage in simulations, but alternative confidence intervals might also be considered such as the complementary log-log transformation [36]. The proposed estimator was used to estimate incidence of relapse by genotype in a cohort of childhood ALL patients, where the genotype was ascertained only on a subsample of the cohort chosen by an optimal sampling approach based on relapse as the event of interest. Interestingly, we can also analyze the incidence of the competing event (toxic death) or of the combined endpoint (relapse or toxic death), but the efficiency could be lower unless the subsampling is adapted to this new endpoint by including a further strata on toxic death in the sampling process. This is particularly important since toxic death is a rare event in this context. We should also remember to avoid random sampling when the event of interest is rare, as discussed in the Simulations section.

In the second case, we dealt with a missing data problem, in which the outcome itself was not available for everybody, since some patients were lost to follow-up. A subsample of lost patients had been tracked to ascertain the outcome, but if this tracking was not possible, a more basic/naïve approach to deal with this informative censoring could have been to identify the variables affecting missingness, post-stratify the sample in homogeneous strata and use the missing probabilities for each strata as sort of sampling weights to adjust the incidence estimate. This approach would make an important assumption of missing at random that might be not appropriate and cannot be tested [7, 15, 21, 37]. However this underlines

how the proposed estimator could be applied also in the presence of missing data.

The code to compute this estimate has been developed in R software [31] under the `survey` package [32] and is available at [33]. The survey package is a flexible package for complex surveys including also two-phase studies. It provides flexible functions to describe the design of the study and to derive sampling fraction accordingly. The package includes functions to estimate survival and to perform a weighted Cox model with standard error properly adjusted for the design and with the possibility to use general weights, as calibrated weights. Our function takes advantage of the facilities of the package (see an example of use in Genotype ascertained on a subset of a clinical trial cohort).

In order to recover the representativeness of the subcohort (phase II) for the entire cohort, we used weights related to the inverse of the probability to be sampled, similarly to the weights of Barlow for case-cohort studies [38]. More general weights can be used, such as calibration weights [7, 39, 40]. The use of calibration weights is advantageous when there is availability of phase I variables that are strongly related to the additional variables ascertained in phase II. This would provide results more representative of phase I data and increase precision. When phase II variables are common genetic polymorphisms, as in our first example, it is unlikely to find any strong relation between phase I and II variables, therefore no big advantage would be expected by calibration.

The estimator can also be extended to a situation where an individual may move among a finite number of states to estimate the Aalen-Johansen probabilities of transition among each state in a multistate framework [41] in the presence of general sampling design.

In order to derive a model-based estimate of incidence (adjusted for possible covariates) two main approaches have been followed in the context of competing risk, the first one based on the cause-specific hazard inspired by Benichou and Gail [16, 42, 43] and the other one based on the subdistribution hazard [19, 44]. The crude cumulative estimator developed by Kovalchik and Pfeiffer [45] for two-phase studies for finite population follows the first approach, and the Cox model for two-phase designs [14] could be used to extend it for infinite population. Under this model we can estimate the effect of a covariate on the cause-specific hazard to address its impact on the event by an aetiological point of view. However it is well known that this does not reflect the impact of the variable on the crude cumulative incidence. The latter effect, even if affected by the incidence of the other competing events, could still be of interest for a public health prospective. Future work will concern the development of a regression model to assess the effect of a covariate on the crude cumulative incidence. The Fine and Gray regression model [19] could be extended to complex sampling by weighting the estimating function of the parameter of interest and working out their influence function.

A Appendix

A.1 Derivation of the risk set for T_k^*

The risk set for the usual survival time T at s is commonly obtained in standard analysis by counting the observed times greater then s. It can be also written as:

$$\hat{Y}_{\cdot}(s) = \hat{P}(T > s^-)\hat{P}(C > s^-) = \hat{Y}_{\cdot}(0)\hat{S}(s^-)\hat{G}(s^-) \quad (12)$$

where $\hat{G}(s)$ is the probability to be free of censoring up to s and is estimated considering censored observations as events and viceversa according to (3). This can proved also in the case of a two-phase design by the following:

$$\begin{aligned}
\hat{Y}_{\cdot}(0)\hat{S}(s^-)\cdot\hat{G}(s^-) &= \hat{Y}_{\cdot}(0)\prod_{u<s}\left[1-\frac{\hat{N}_{\cdot\cdot}(du)}{\hat{Y}_{\cdot}(u)}\right]\cdot \\
&\quad \prod_{u<s}\left[1-\frac{\hat{N}^c_{\cdot}(du)}{\hat{Y}_{\cdot}(u)-\hat{N}_{\cdot\cdot}(du)}\right] = \\
&= \hat{Y}_{\cdot}(0)\prod_{u<s}\left[\frac{\hat{Y}_{\cdot}(u)-[\hat{N}_{\cdot\cdot}(du)+\hat{N}^c_{\cdot}(du)]}{\hat{Y}_{\cdot}(u)}\right] = \\
&= \hat{Y}_{\cdot}(0)\prod_{u<s}\frac{\hat{Y}_{\cdot}(u^+)}{\hat{Y}_{\cdot}(u)} = \frac{\hat{Y}_{\cdot}(0)\hat{Y}_{\cdot}(s)}{\hat{Y}_{\cdot}(0)} = \\
&= \sum_{i=1}^{N}[\xi_i Y_i(s)/\pi_i]
\end{aligned} \quad (13)$$

where $\hat{N}^c_{\cdot}(t) = \sum_{i=1}^{N}[\xi_i I(X_i \le t, \Delta_i = 0)/\pi_i]$ denotes the number of censoring up to time t and $\hat{N}_{\cdot\cdot}(t) = \sum_{i=1}^{n}\hat{N}_{i\cdot}(t)$ the total count of events observed up to time t.

The derivation of the risk set for T_k^* at s can be obtained by:

$$\begin{aligned}
\hat{Y}^*_{\cdot k}(s) &= \hat{P}\left(T_k^* > s^-\right)\hat{P}(C > s^-) = \hat{Y}_{\cdot}(0)\cdot\left[1-\hat{F}_k(s^-)\right]\hat{G}(s^-) = \\
&= \hat{Y}_{\cdot}(0)\cdot\left[\hat{S}(s^-)+\sum_{l\neq k}\hat{F}_l(s^-)\right]\cdot\hat{G}(s^-) = \\
&= \hat{Y}_{\cdot}(s)+\hat{Y}_{\cdot}(0)\cdot\sum_{l\neq k}\hat{F}_l(s^-)\cdot\hat{G}(s^-)
\end{aligned} \quad (14)$$

By writing $\hat{F}_l(s^-)$ as empirical cumulative distribution function $\hat{F}_l(s^-) = \frac{1}{\hat{Y}_{\cdot}(0)}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\frac{I(X_i\le s^-;\varepsilon_i=l)}{\hat{G}(X_i^-\wedge s^-)}$ [25, 46], the risk set becomes:

$$\hat{Y}^*_{\cdot k}(s) = \hat{Y}_{\cdot}(s) + \sum_{l\neq k}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i} I(X_i \leq s^-; \varepsilon_i = l)\frac{\hat{G}(s^-)}{\hat{G}\left(X_i^- \wedge s^-\right)} \tag{15}$$

that can be simplified to

$$\sum_{i=1}^{N}\frac{\xi_i}{\pi_i} Y_i(s) + \sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\left[\sum_{l\neq k} N_{il}(s^-)\cdot \hat{G}\left(s^-|X_i^-\right)\right] \tag{16}$$

that is equivalent to (4).

The first summation estimates the usual total number of subjects at risk at s, where the condition $X_i = \min(T_i, C_i) \geq s$ is satisfied. This in fact implies $\min(T^*_{ki}, C_i) \geq s$, i.e. being at risk for T implies being also at risk for T^*_k. The second summation estimates the number of subjects who had other events before s, satisfying the condition $X_i = \min(T_i, C_i) < s$, $\Delta_i = 1$ and $\varepsilon_i \neq k$ which implies $T^*_{ki} = \infty > s$ and completes the number at risk at s for T^*_k. While the first part is exposed to censoring, the contribute of each subject observed to fail of cause $l \neq k$, $\sum_{l\neq k} N_{il}(s^-)$, would remain equal to 1 up to ∞, thus ignoring possible censoring, given that C_i is (usually) not observable if $T_i < C_i$. A possible way to deal with this inconsistency is to mimic the presence of random censoring acting on the infinite times, by weighting the unitary contributions $\sum_{l\neq k} N_{il}(s^-)$ by the estimate $\hat{G}(s^-|X_i^-)$ of $P(C > s^-|C > X_i^-) = G(s^-|X_i)$, where $G(t) = P(C > t)$ is estimated by $\hat{G}(t) = \prod_{s\leq t}\left[1 - \frac{\sum_{i=1}^N \xi_i N_i^c(ds)/\pi_i}{\sum_{i=1}^N \xi_i Y_i(s)/\pi_i}\right]$, with $N_i^c(s) = I(X_i \leq s, \Delta_i = 0)$. This weight assumes value 1 before X_i and decreases afterword according to the censoring distribution.

Of note, an alternative expression for $\hat{Y}^*_{\cdot k}(s)$ derives substituting $\hat{Y}_{\cdot}(0)\hat{G}(s^-) = \frac{\hat{Y}_{\cdot}(s)}{\hat{S}(s^-)}$ from (12) in (14):

$$\hat{Y}^*_{\cdot k}(s) = \hat{Y}_{\cdot}(0)\hat{G}(s^-)\cdot\left[1 - \hat{F}_k(s^-)\right] = \hat{Y}_{\cdot}(s)\frac{\left[1 - \hat{F}_k(s^-)\right]}{\hat{S}(s^-)} \tag{17}$$

This shows as $\hat{Y}_{\cdot}(s)$ is upweighted by a multiplier that gets greater as the action of competing events gets larger, accounting for the fact that subjects that experienced events of type $l \neq k$ will never experience event k as first.

A.2 Proof of equivalence (7)

The equality between the cumulative incidence of the artificial variable T^* in (6) and the Aalen-Johansen type estimator (that for the purpose of this proof will be denoted as $\hat{F}^{AJ}_k(t)$) holds true if and only if, $\forall t$:

$$\hat{F}^{AJ}_k(t) = \int_0^t \hat{S}(s^-)\hat{\Lambda}_k(ds) = 1 - \prod_{s\leq t}\left[1 - \hat{\Lambda}^*_k(ds)\right] = \hat{F}_k(t) \tag{18}$$

which can be proved by induction. Both quantities are step functions, changing value at each occurrence of type k events. At the time t where the first event of type k is observed, $\hat{\Lambda}_k(dt) = \hat{\Lambda}^*_k(dt)$ being from (4) $\hat{Y}_{\cdot}(t) = \hat{Y}^*_{\cdot k}(t)$. If this was the first event overall, then $\hat{S}(0) = 1$ and Eq. 18 is satisfied, otherwise $\hat{F}^{AJ}_k(t) = [1 - 1/\hat{Y}_{\cdot}(0)]\cdot 1/[\hat{Y}_{\cdot}(0) - 1] = 1/\hat{Y}_{\cdot}(0) = 1 - [1 - 1/\hat{Y}_{\cdot}(0)] = \hat{F}_k(t)$.

Now, assuming that (18) holds true for a given t^-, this implies $\hat{F}^{AJ}_k(t) = \hat{F}_k(t)$ if and only if, from (18):

$$\hat{F}^{AJ}_k(t^-) + \hat{S}(t^-)\hat{\Lambda}_k(dt) = 1 - (1 - \hat{F}_k(t^-))\left(1 - \hat{\Lambda}^*_k(dt)\right)$$

and using the equality at t^-:

$$\begin{aligned}
\hat{F}_k(t^-) + \hat{S}(t^-)\hat{\Lambda}_k(dt) &= \hat{F}_k(t^-) + \left[1 - \hat{F}_k(t^-)\right]\hat{\Lambda}^*_k(dt)\\
\hat{S}(t^-)\hat{\Lambda}_k(dt) &= \left[1 - \hat{F}_k(t^-)\right]\hat{\Lambda}^*_k(dt)\\
\hat{\Lambda}_k(dt) &= \frac{1 - \hat{F}_k(t^-)}{\hat{S}(t^-)}\hat{\Lambda}^*_k(dt)\\
\frac{\hat{N}_k(dt)}{\hat{Y}_{\cdot}(t)} &= \frac{1 - \hat{F}_k(t^-)}{\hat{S}(t^-)}\frac{\hat{N}_k(dt)}{\hat{Y}^*_{\cdot k}(t)}\\
\hat{Y}^*_{\cdot k}(t) &= \hat{Y}_{\cdot}(t)\cdot\frac{1 - \hat{F}_k(t^-)}{\hat{S}(t^-)}
\end{aligned}$$

That is proved by (17).

A.3 Weak convergence of $\hat{\Lambda}^*_k(t)$

Lin showed that the normalised Horvitz-Thompson estimators of the number of events $\sqrt{N}\left[\hat{N}_{\cdot}(t) - N_{\cdot}(t)\right]$, number at risk $\sqrt{N}\left[\hat{Y}_{\cdot}(t) - Y_{\cdot}(t)\right]$, cumulative hazard $\sqrt{N}[\hat{\Lambda}(t) - \Lambda(t)]$ and survival function $\sqrt{N}\left[\hat{S}(t) - S(t)\right]$ (and analogously $\sqrt{N}\left[\hat{G}(t) - G(t)\right]$) are asymptotically multivariate zero-mean normal [14]. Firstly, we concentrate on the normalised Horvitz-Thompson estimators of the modified at risk process: $\sqrt{N}\left[\hat{Y}^*_{\cdot k}(t) - Y^*_{\cdot k}(t)\right] = \sqrt{N}\sum_{i=1}^N \frac{\xi_i - \pi_i}{\pi_i} Y^*_{\cdot k}(t) = \sqrt{N}\sum_{i=1}^N \frac{\xi_i - \pi_i}{\pi_i} Y_{\cdot k}(t) + \sqrt{N}\sum_{i=1}^N \frac{\xi_i - \pi_i}{\pi_i}\sum_{l\neq k} N_{il}(t^-)\hat{G}(t^-|X_i)$. The first term represents the estimator of the number at risk, that Lin showed to be asymptotically multivariate zero-mean normal [14]. We concentrate now on the second term:

$$\sqrt{N}\sum_{i=1}^{N}\left[\frac{\xi_i}{\pi_i}\cdot\sum_{l\neq k}N_{il}(t^-)\hat{G}(t^-|X_i)-\sum_{l\neq k}N_{il}(t^-)\hat{G}(t^-|X_i)\right]=$$
$$=\sqrt{N}\sum_{i=1}^{N}\left[\hat{G}(t^-|X_i)\left\{\frac{\xi_i}{\pi_i}\sum_{l\neq k}N_{il}(t^-)-\sum_{l\neq k}N_{il}(t^-)\right\}\right]=$$
$$=\sqrt{N}\left[\hat{G}(t^-|X_i)\left\{\sum_{l\neq k}\hat{N}_{\cdot l}(t^-)-\sum_{l\neq k}\sum_{i=1}^{N}N_{il}(t^-)\right\}\right]=$$
$$=\sqrt{N}\int_0^{t^-}\hat{G}(s^-|X_i)\left\{\sum_{l\neq k}\hat{N}_{\cdot l}(ds-)-\sum_{l\neq k}\sum_{i=1}^{N}N_{il}(ds^-)\right\}=$$
$$=\sqrt{N}\int_0^{t^-}G(t^-|X_i)\left\{\sum_{l\neq k}\hat{N}_{\cdot l}(ds-)-\sum_{l\neq k}\sum_{i=1}^{N}N_{il}(ds^-)\right\}+$$
$$+o_p(1)$$

that using Lemma 1 in [14] also converges to a zero-mean Gaussian process.

We want to prove that also $\sqrt{N}\left[\hat{\Lambda}_k^*(t)-\Lambda_k^*(t)\right]=\sqrt{N}\left[\tilde{\Lambda}_k^*(t)-\Lambda_k^*(t)\right]+\sqrt{N}\left[\hat{\Lambda}_k^*(t)-\tilde{\Lambda}_k^*(t)\right]$ converges to a zero-mean normal, where $\tilde{\Lambda}_k^*(t)$ represents the crude cumulative incidence estimator that we would have obtained if complete information $(X_i, \Delta_i\varepsilon_i, Z_i)$ was known for all the subjects in phase I sample $(i = 1\ldots N)$ [18]. Fine and Gray proved that the first term converges weakly to a zero-mean Gaussian process [19].

The second term results:

$$\sqrt{N}\left[\hat{\Lambda}_k^*(t)-\tilde{\Lambda}_k^*(t)\right]=\sqrt{N}\left[\int_0^t\frac{\hat{N}_k(ds)}{\hat{Y}_{\cdot k}^*(s)}-\int_0^t\frac{N_k(ds)}{Y_{\cdot k}^*(s)}\right]=$$
$$=\sqrt{N}\left[\int_0^t\frac{\hat{N}_k(ds)}{\hat{Y}_{\cdot k}^*(s)}-\int_0^t\frac{N_k(ds)}{Y_{\cdot k}^*(s)}+\int_0^t\frac{N_k(ds)}{\hat{Y}_{\cdot k}^*(s)}-\int_0^t\frac{N_k(ds)}{\hat{Y}_{\cdot k}^*(s)}\right]=$$
$$=\sqrt{N}\left[\int_0^t\frac{\hat{N}_k(ds)-N_k(ds)}{\hat{Y}_{\cdot k}^*(s)}-\int_0^t\frac{N_k(ds)\left[\hat{Y}_{\cdot k}^*(s)-Y_{\cdot k}^*(s)\right]}{\hat{Y}_{\cdot k}^*(s)Y_{\cdot k}^*(s)}\right].$$

It then follows that also $\sqrt{N}\left[\hat{\Lambda}_k^*(t)-\tilde{\Lambda}_k^*(t)\right]$ converges weakly to a zero-mean Gaussian process.

A.4 Influence function for $\hat{\Lambda}_k^*(t)$

The estimator $\hat{\Lambda}_k^*(t)$ can be expressed as a differentiable function g of the estimated total number of events of type k and total number at risk for T^* up to t:

$$\hat{\Lambda}_k^*(t)=g(\hat{N}_{\cdot k}(dt),\hat{Y}_{\cdot k}^*(t))=\int_0^t\frac{\hat{N}_{\cdot k}(ds)}{\hat{Y}_{\cdot k}^*(s)}=$$
$$=\int_0^t\frac{\sum_{i=1}^{N}\xi_i\cdot N_{ik}(ds)w_i}{\sum_{i=1}^{N}\xi_i\cdot Y_{ik}^*(s)w_i}\tag{19}$$

where $w_i = 1/\pi_i$ and ξ_i indicates whether subject i is withdrawn in the phase II sample.

The difference between the true and estimated cumulative hazard can be expressed as a sum of influence functions: $\hat{\Lambda}_k^*(t)-\Lambda_k^*(t)=\sum_{i=1}^{N}z_{ik}^*+o(1/\sqrt{N})$, where z_{ik}^* is the influence function of the i^{th} subject. Demnati and Rao [27] proved that we can express the influence function of subject i as $z_{ik}^*(t)=\frac{\partial g\left(\hat{N}_{\cdot k}(dt),\hat{Y}_{\cdot k}^*(t)\right)}{\partial w_i}$. The influence function of $\hat{\Lambda}_k^*(t)$ of the i^{th} subject can thus be derived as:

$$z_{ik}^*(t)=\int_0^t\frac{N_{ik}(ds)\hat{Y}_{\cdot k}^*(s)-\frac{\partial\hat{Y}_{\cdot k}^*(s)}{\partial w_i}\hat{N}_{\cdot k}(ds)}{\hat{Y}_{\cdot k}^*(s)^2}\tag{20}$$

Being $\hat{Y}_{\cdot k}^*(s)=\sum_{i=1}^{N}\xi_i w_i\cdot\left[Y_i(s)+\frac{I(X_i\leq s^-;\varepsilon_i\neq k)\hat{G}(s^-)}{\hat{G}(X_i^-\wedge s^-)}\right]$ and being $\hat{G}(s)=\prod_{u\leq s}\left[1-\frac{\sum_{i=1}^{N}\xi_i N_i^c(du)/\pi_i}{\sum_{i=1}^{N}\xi_i Y_i(u)/\pi_i}\right]$, the derivative of $\hat{Y}_{\cdot k}^*(s)$ with respect to w_i results:

$$\frac{\partial\hat{Y}_{\cdot k}^*(s)}{\partial w_i}=\frac{\partial\left\{\xi_i w_i\cdot\left[Y_i(s)+\frac{I(X_i\leq s^-;\varepsilon_i\neq k)\hat{G}(s^-)}{\hat{G}(X_i^-\wedge s^-)}\right]+\sum_{j\neq i}\xi_j w_j\cdot\left[Y_j(s)+\frac{I(X_j\leq s^-;\varepsilon_j\neq k)\hat{G}(s^-)}{\hat{G}\left(X_j^-\wedge s^-\right)}\right]\right\}}{\partial w_i}=$$
$$=\xi_i Y_i(s)+\xi_i\frac{\partial\left[w_i\frac{I(X_i\leq s^-;\varepsilon_i\neq k)\hat{G}(s^-)}{\hat{G}(X_i^-\wedge s^-)}\right]}{\partial w_i}+\frac{\partial\left[\sum_{j\neq i}\frac{\xi_j w_j I(X_j\leq s^-;\varepsilon_j\neq k)\hat{G}(s^-)}{\hat{G}\left(X_j^-\wedge s^-\right)}\right]}{\partial w_i}=\tag{21}$$
$$=\xi_i Y_i(s)+\xi_i\frac{I(X_i\leq s^-;\varepsilon_i\neq k)\hat{G}(s^-)}{\hat{G}\left(X_i^-\wedge s^-\right)}+\sum_{j=1}^{N}\xi_j w_j I(X_j\leq s^-;\varepsilon_j\neq k)\frac{\partial}{\partial w_i}\left[\frac{\hat{G}(s^-)}{\hat{G}\left(X_j^-\wedge s^-\right)}\right].$$

Please note that the last addendum accounts for the fact that $\hat{G}(t)$ is estimated using information of subject i. For $v \leq u$:

$$\frac{\partial}{\partial w_i}\left[\frac{\hat{G}(u)}{\hat{G}(v)}\right] = \frac{\hat{G}(u)\int_0^u \frac{dN_i^c(s)-\hat{\lambda}^c(s)Y_i(s)}{\hat{Y}_{\cdot}(s)}\hat{G}(v) - \hat{G}(v)\int_0^v \frac{dN_i^c(s)-\hat{\lambda}^c(s)Y_i(s)}{\hat{Y}_{\cdot}(s)}\hat{G}(u)}{\hat{G}(v)^2} = \\ = \frac{\hat{G}(u)\hat{G}(v)\int_{v^+}^u \frac{dN_i^c(s)-\hat{\lambda}^c(s)Y_i(s)}{\hat{Y}_{\cdot}(s)}}{\hat{G}(v)^2} = \frac{\hat{G}(u)\int_{v^+}^u \frac{dN_i^c(s)-\hat{\lambda}^c(s)Y_i(s)}{\hat{Y}_{\cdot}(s)}}{\hat{G}(v)} \tag{22}$$

where the superscript c indicates the quantities related to the censoring process, i.e. $N_i^c(u) = I(X_i \leq u, \Delta_i = 0)$ is the indicator of censoring for subject i up to time u and $\hat{\lambda}^c(u) = \hat{N}_{\cdot}^c(du)/\hat{Y}_{\cdot}(u)$ the instantaneous hazard of censoring. The derivative of $\hat{Y}_{\cdot k}^*(u)$ becomes:

$$\frac{\partial \hat{Y}_{\cdot k}^*(s)}{\partial w_i} = \\ Y_i(s) + I(X_i \leq s^-; \varepsilon_i \neq k) \cdot \frac{\hat{G}(s^-)}{\hat{G}\left(X_i^- \wedge s^-\right)} + \sum_{j=1}^N \xi_j w_j I(X_j \leq s^-; \varepsilon_j \neq k) \frac{\hat{G}(s^-)}{\hat{G}\left(X_j^- \wedge s^-\right)} \left[\int_{X_j}^{s^-} \frac{N_i^c(du) - \hat{\lambda}^c(u)Y_i(u)}{\hat{Y}_{\cdot}(u)}\right] \tag{23}$$

Thus, the influence function of $\hat{\Lambda}_k^*(t)$ of the i^{th} subject results:

$$z_{ik}^*(t) = \int_0^t \frac{N_{ik}(ds)\hat{Y}_{\cdot k}^*(s) - \hat{N}_{\cdot k}(ds)\left[Y_i(s) + \frac{I(X_i \leq s^-; \varepsilon_i \neq k)\hat{G}(s^-)}{\hat{G}(X_i^- \wedge s^-)} + \sum_{j=1}^N \frac{\xi_j w_j I(X_j \leq s^-; \varepsilon_j \neq k)\hat{G}(s^-)}{\hat{G}\left(X_j^- \wedge s^-\right)} \int_{X_j}^{s^-} \frac{N_i^c(du) - \hat{\lambda}^c(u)Y_i(u)}{\hat{Y}_{\cdot}(s)}\right]}{\hat{Y}_{\cdot k}^*(s)^2} = \\ = \int_0^t \frac{N_{ik}(ds) - \hat{\Lambda}_k^*(ds)\left[Y_i(s) + \frac{I(X_i \leq s^-; \varepsilon_i \neq k)\hat{G}(s^-)}{\hat{G}(X_i^- \wedge s^-)} + \sum_{j=1}^N \frac{\xi_j I(X_j \leq s^-; \varepsilon_j \neq k)\hat{G}(s^-)}{\pi_j \hat{G}\left(X_j^- \wedge s^-\right)} \int_{X_j}^{s^-} \frac{N_i^c(du) - \hat{\lambda}^c(u)Y_i(u)}{\hat{Y}_{\cdot}(u)}\right]}{\hat{Y}_{\cdot k}^*(s)} \tag{24}$$

where $w_j = 1/\pi_j$. By defining $M_i^c(s,t) = \int_s^t \frac{N_i^c(du) - \hat{\lambda}^c(u)Y_i(u)}{\hat{Y}_{\cdot}(u)}$ as the influence function of subject i on censoring, $v_j = \frac{\xi_j}{\pi_j}\sum_{l \neq k} N_{jl}(s^-)\hat{G}(s^-|X_j^-)$, and $M_{ik}(s) = \left[N_{ik}(s) - \hat{\Lambda}_k^*(s)Y_{ik}^*(s)\right]/\hat{Y}_{\cdot k}^*(s)$.

Thus it can be reduced as:

$$z_{ik}^*(t) = \int_0^t \frac{N_{ik}(ds) - \hat{\Lambda}_k^*(ds)Y_{ik}^*(s) - \hat{\Lambda}_k^*(ds)\left[\sum_{j=1}^N \frac{\xi_j}{\pi_j}\sum_{l \neq k} N_{jl}(s^-)\hat{G}\left(s^-|X_j^-\right) M_i^c\left(X_j, s^-\right)\right]}{\hat{Y}_{\cdot k}^*(s)} = \\ = \int_0^t M_{ik}(ds) - \frac{\hat{\Lambda}_k^*(ds)\left[\sum_{j=1}^N v_j M_i^c\left(X_j, s^-\right)\right]}{\hat{Y}_{\cdot k}^*(s)}$$

Abbreviations
ALL: Acute lymphoblastic leukemia; ART: Antiretroviral therapy; BM: Bone-marrow; CI: Confidence interval; GST-T1: Glutathione S-transferase-θ; HR: Hazard ratio.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
PR performed the theoretical derivation, analyses and simulation studies and drafted the manuscript. LA, DVG, MGV contributed to the analyses and to the manuscript. All authors read and approved the final manuscript.

Acknowledgements
The authors thank Thomas Lumley for useful comments, Raffaella Franca and Marco Rabusin for the ALL data and Jeff Martin and Elvin Geng for the HIV data. PR was supported by the grant SIR RBSI14LOVD of the Italian Ministry of Education, University and Research. This research was partially supported by AIRC (2013-14634, MGV) and by NIH (U01 AI069911).

Author details
[1]Center of Biostatistics for Clinical Epidemiology, School of Medicine and Surgery, University of Milano-Bicocca, via Cadore 48, 20900 Monza, Italy. [2]Department of Epidemiology and Biostatistics, University of California, San Francisco, California.

Received: 17 September 2015 Accepted: 17 December 2015
Published online: 11 January 2016

References

1. Neyman J. Contribution to the theory of sampling human populations. J Am Stat Assoc. 1938;33(201):101–16.
2. Borgan Ø, Samuelsen SO. A review of cohort sampling designs for Cox's regression model: Potentials in epidemiology. Norsk Epidemiol. 2003;13: 239–48.
3. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986;73(1):1–11.
4. Samuelsen SO. A psudolikelihood approach to analysis of nested case-control studies. Biometrika. 1997;84(2):379–94.
5. Langholz B, Borgan Ø. Counter-matching: a stratified nested case-control sampling method. Biometrika. 1995;82(1):69–79.
6. Rudolph KE, Gary SW, Stuart EA, Glass TA, Marques AH, Duncko R, et al. The association between cortisol and neighborhood disadvantage in a US population-based sample of adolescents. Health Place. 2014;25:68–77.
7. Lumley TS. Complex Surveys: A Guide to Analysis Using R. 1st ed. Inc JWS, editor. Wiley Series in Survey Methodology. Hoboken, New Jersey: John Wiley & Sons; 2010.
8. Breslow N, Lumley T, Ballantyne C, Chambless L, Kulich M. Using the Whole Cohort in the Analysis of Case-Cohort Data. Am J Epidemiol. 2009;169(11):1398–405.
9. Anderson GL, Manson J, Wallace R, Lund B, Hall D, Davis S, et al. Implementation of the women's health initiative study design. Ann Epidemiol. 2003;13(9, Supplement):S5—17.
10. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, et al. The Cardiovascular Health Study: design and rationale. Ann Epidemiol. 1991;1(3):263–76.
11. Franca R, Rebora P, Basso G, Biondi A, Cazzaniga G, Crovella S, et al. Glutathione S-transferase homozygous deletions and relapse in childhood acute lymphoblastic leukemia: a novel study design in a large Italian AIEOP cohort. Pharmacogenomics. 2012;13(16):1905–16.
12. Geng E, Emenyonu N, Bwana M, Glidden D, Martin J. Sampling-based approach to determining outcomes of patients lost to follow-up in antiretroviral therapy scale-up programs in Africa. JAMA. 2008;300(5): 506–7. Available from: http://dx.doi.org/10.1001/jama.300.5.506.
13. Rebora P, Valsecchi MG. Survival estimation in two-phase cohort studies with application to biomarkers evaluation. Stat Methods Med Res. 2014. in press. doi:10.1177/0962280214534411.
14. Lin DY. On fitting Cox's proportional hazards models to survey data. Biometrika. 2000;87(1):37–47.
15. Frangakis CE, Rubin DB. Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring. Biometrics. 2001;57(2):333–42.
16. Borgan Ø. Estimation of covariate-dependent Markov transition probabilities from nested case-control data. Stat Methods Med Res. 2002;11(2):183–202.
17. Aalen OO, Borgan Ø, Fekjær H. Covariate adjustment of event histories estimated from Markov chains: the additive approach. Biometrics. 2001;57(4):993–1001.
18. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. The Annals of Statistics. 1988;16(3):1141–54.
19. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999;94(446):496–509.
20. Antolini L, Biganzoli EM, Boracchi P. Crude cumulative incidence in the form of a Horvitz-Thompson like and Kaplan-Meier like estimator. COBRA Preprint Series. 2006. 10 http://biostats.bepress.com/cobra/art10.
21. Särndal C, Swensson B. A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. Int Stat Rev. 1987;55(3):279–94.
22. Kang S, Cai J. Marginal hazards model for case-cohort studies with multiple disease outcomes. Biometrika. 2009;96(4):887–901.
23. Marubini E, Valsecchi MG. Analysing survival data from clinical trials and observational studies. Chichester, England: Wiley-Interscience; 2004.
24. Bernasconi D, Antolini L. Description of survival data extended to the case of competing risks: a teaching approach based on frequency tables. Epidemiol Biostat Public Heal. 2013;11(1):. e8874–1:e8874–10.
25. Satten GA, Datta S. The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. Am Stat. 2001;55(3): 207–10.
26. Breslow NE, Wellner JA. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. Scand J Stat. 2007;34(1):86–102.
27. Demnati A, Rao JNK. Linearization variance estimators for model parameters from complex survey data. Surv Methodol. 2010;36:193–201.
28. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. Stat Biosci. 2009;1(1):32–49.
29. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952;47(260):663–85.
30. de Wreede LC, Fiocco M, Putter H. mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. J Stat Softw. 2011;38(7):1–30. Available from: http://www.jstatsoft.org/v38/i07/.
31. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. Available from: http://www.R-project.org.
32. Lumley T. Analysis of complex survey samples. J Stat Softw. 2004;9(8):1–19.
33. Rebora P. R code to estimate crude incidence in two-phase designs. 2015. Accessed: 2015-10-27. http://dx.doi.org/10.6070/H4F18WRG.
34. Tsiatis AA. A nonidentifiability aspect of the problem of competing risks. Proc Natl Acad Sci U S A. 1975;72(1):20–2.
35. Geng EH, Glidden DV, Bwana MB, Musinguzi N, Emenyonu N, Muyindike W, et al. Retention in care and connection to care among HIV-infected patients on antiretroviral therapy in Africa: estimation via a sampling-based approach. PLoS One. 2011;e21797:7.
36. Glidden DV. Robust inference for event probabilities with Non-Markov event data. Biometrics. 2002;58(2):361–68.
37. Lin DY, Ying Z. Cox regression with incomplete covariate measurements. J Am Stat Assoc. 1993;88:1341–9.
38. Barlow W, Ichikawa L, Rosner D, Izum iS. Analysis of case-cohort designs. J Clin Epidemiol. 1999;52(12):1165–72.
39. Scott AJ, Wild CJ. Fitting regression models with response-biased samples. Can J Stat. 2011;39(3):519–36.
40. Rudolph KE, Diaz I, Rosenblum M, Stuart EA. Estimating population treatment effects from a survey sub-sample. Am J Epidemiol. 2014;180: 737–48.
41. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scand J Stat. 1978;5(3):141–50. Available from: http://www.jstor.org/stable/4615704.

42. Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. Biometrics. 1990;46(3):813–26. Available from: http://www.jstor.org/stable/2532098.
43. Langholz B, Borgan Ø. Estimation of absolute risk from nested case-control data. Biometrics. 1997;53(2):767–74.
44. Wolkewitz M, Cooper BS, Palomar-Martinez M, Olaechea-Astigarraga P, Alvarez-Lerma F, Schumacher M. Nested case-control studies in cohorts with competing events. Epidemiology. 2014;25(1):122–5.
45. Kovalchik S, Pfeiffer R. Population-based absolute risk estimation with survey data. Lifetime Data Anal. 2014;20(2):252–75.
46. Jewell NP, Lei X, Ghani AC, Donnelly CA, Leung GM, Ho LM, et al. Non-parametric estimation of the case fatality ratio with competing risks data: an application to Severe Acute Respiratory Syndrome (SARS). Stat Med. 2007;26(9):1982–98.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit