SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2016, Vol. 25(6) 2895–2908
© The Author(s) 2014
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280214534411
smm.sagepub.com
SAGE

Survival estimation in two-phase cohort studies with application to biomarkers evaluation

Paola Rebora and Maria Grazia Valsecchi

Abstract

Two-phase studies are attractive for their economy and efficiency in research settings where large cohorts are available for investigating the prognostic and predictive role of novel genetic and biological factors. In this type of study, information on novel factors is collected only in a convenient subcohort (phase II) drawn from the cohort (phase I) according to a given (optimal) sampling strategy. Estimation of survival in the subcohort needs to account for the design. The Kaplan–Meier method, based on counts of events and of subjects at risk in time, must be applied accounting, with suitable weights, for the sampling probabilities of the subjects in phase II, in order to recover the representativeness of the subcohort for the entire cohort. The authors derived a proper variance estimator of survival by linearization. The proposed method is applied in the context of a two-phase study on childhood acute lymphoblastic leukemia, which was planned in order to evaluate the role of genetic polymorphisms on treatment failure due to relapse. The method has shown satisfactory performance through simulations under different scenarios, including the case–control setting, and proved to be useful for describing results in the clinical example.

Keywords

two-phase design, survival, optimal sampling, case–control, missing data

1 Introduction

Longitudinal cohort studies provide the ideal setting to identify new biomarkers (as risk factors or predictors of response to treatment), since many of such studies have stored biologic samples from thousands of individuals who are being followed up over many years. However, the combination between large cohorts and expensive new technologies makes infeasible to measure novel biomarkers on the entire cohort and efficient study designs are needed. The traditional approach consists in the nested case–control design, where a sample is artificially selected to include all cases with a given event or disease (especially if rare) and a subset of controls that are matched to the cases for possible confounders. Although this design is efficient, it also presents limitations: by the standard analysis, one can only evaluate risk/protective factors for the event type that defines the cases (and not for

Center of Biostatistics for Clinical Epidemiology, Department of Health Sciences, University of Milano-Bicocca, Monza, Italy

Corresponding author:
Paola Rebora, Center of Biostatistics for Clinical Epidemiology, Department of Health Sciences, University of Milano-Bicocca, Via Cadore 48, 20900 Monza, Italy.
Email: paola.rebora@unimib.it

other possible events of interest). Only with a weighted/pseudo likelihood approach, it is possible to reuse controls for different end-points,[1,2] but this is rarely done in practice. This considerably limits the potentiality of the study, in particular when precious samples from biobanks are used to measure additional "costly" covariates. Moreover, the traditional statistical analysis includes only the selected sample and does not use other valuable information available for the entire cohort on variables routinely ascertained. A more flexible approach is the case–cohort design, where all the cases are selected together with a subcohort randomly sampled from the full cohort at start of follow-up.[3,4] This design has been extended to allow for stratified sampling (counter-matching).[4,5]

An alternative approach, that can be considered as a sort of generalization of designs above, is to use a two-phase sampling technique[6–8] that considers the entire cohort as the first-phase sample from the population of interest, in the presence of strata. In the second phase, subsamples are drawn from each stratum in the cohort, with different sampling probabilities, and biomarkers are measured only on these individuals. The statistical analysis of the second-phase sample makes use of the information available on the whole cohort and mimics it by using weights related to the sampling probabilities, while analyzing the additional covariates available in the second phase only.[9] Advantages of the two-phase design include the opportunity to use all the second-phase data, without having to comply to a strict matching, the possibility to use different models, and to accommodate different time scales in the analysis. Also, this design is advantageous when exposure is rare, because this can be accounted for, through stratification, in sampling from the cohort.[10] The main advantage of the two-phase design is that the estimation of event incidence and population frequency of risk factors (i.e., genotypes) is straightforward.

The estimate of survival in the cohort (phase I) can be performed with standard Kaplan–Meier (KM) estimate, while in the second-phase sample survival estimation has to account for the design through proper weighting.

The aim of this paper is to develop a weighted KM survival estimator and its variance in the framework of a two-phase design. In the following section, after presenting the motivating example, we introduce a nonparametric estimator and its variance, derived by linearization, under two-phase sampling. The properties of the proposed method are investigated through simulations. An example of two-phase design and application of the proposed estimator to childhood acute lymphoblastic leukemia (ALL) is described and the last section is dedicated to the discussion.

2 Methods

2.1 Motivating context

The clinical context that motivated this work is represented by a study on childhood ALL, performed in order to evaluate the role of different genetic polymorphisms on treatment failure due to relapse. For the purpose of this paper, we will concentrate on the homozygous deletion (or null genotype) in glutathione S-transferase-θ (GST-T1) gene. GST-T1 exhibits a common genetic polymorphism in Caucasians, with 13%–26% of individuals displaying a homozygous deletion of the gene. Subjects carrying the null variants fail to express the GST-T1 enzyme that is involved in drug metabolism. This deletion was proposed as influencing relapse risk in childhood ALL, although conflicting results have been reported.[11] Clinical information was 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 Associazione Italiana di Ematologia Pediatrica (AIEOP) centers between September 2000 and July 2006. For most of these patients, biological samples stored at diagnosis were available, but a parsimonious use of these specimens motivated the choice of a two-phase design.[12] A subsample for the assessment of GST-T1 genotype was chosen after classifying patients into six strata according to the event of interest (relapse/no relapse) and a three-level risk group stratification defined by prognostic features in the treatment protocol. Patients were sampled at random without replacement from the six strata, according to an optimal sampling strategy.[13]

The full cohort of 1999 patients represents the phase I sample, for which standard clinical information are available, while genotype is ascertained in the phase II sample only. The aim is to estimate the impact of GST-T1 deletion on relapse-free interval.

2.2 Survival estimation in two-phase design

Let T_i be the failure time and C_i the censoring time of subject i ($i = 1 \ldots M$) in a cohort (phase I random sample) of size M. The minimum between T_i and C_i ($Z_i = \min(T_i, C_i)$) is observed, and $\Delta_i = I(T_i \leq C_i)$ indicates whether the survival time or the censoring time is observed, as commonly done in survival analysis. T_i and C_i are assumed to be independent. The right continuous counting process $dN_i(t)$ records the number of events experienced at time t (i.e., $dN_i(t) = I(Z_i = t) \cdot \Delta_i$), and $Y_i(t)$ indicates whether the subject is at risk and under observation at time t (at-risk process $Y_i(t) = I(Z_i \geq t)$). By defining the counting process $N_i(t) = \int_0^t dN_i(u)$, the sum over the sample $N.(t) = \sum_{i=1}^{M} N_i(t)$ is the total number of events up to time t, and $Y.(t) = \sum_{i=1}^{M} Y_i(t)$ is the total number of subjects at risk at time t.

Suppose that a biomarker (or additional information) is measured on a subset $m < M$ of subjects drawn according to a certain sampling design (phase II). The goal is to estimate the survival function $S(t) = P(T > t)$, where T is the failure time random variable, in subgroups defined according to the variable ascertained only in the phase II sample. Under random sampling, the KM estimator could be used on the phase II sample leading to an unbiased estimate. However, under efficient sampling adopted in phase II, which samples with higher probability in more informative strata, a biased estimate would be obtained. Borrowing results from the survey theory, unbiased estimates of the total number of events and persons at risk that one would observe on the entire cohort can be obtained using an Horvitz–Thompson approach. These estimated counts can then be used to calculate an unbiased KM estimate as follows.

The total number of events $N.(t)$ and total number of persons at risk $Y.(t)$ at time t can be estimated from the phase II sample (accounting for the sampling design) by

$$d\hat{N}.(t) = \sum_{i=1}^{m} dN_i(t)/\pi_i \quad \text{and} \quad \hat{Y}.(t) = \sum_{i=1}^{m} Y_i(t)/\pi_i \tag{1}$$

where π_i is the probability of the sampling unit i to be selected for phase II from the cohort (i.e., conditional on the realized random phase I sample). The representativeness of the subcohort is recovered by using $1/\pi_i$ as weights.

In our example, 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 is higher for strata that include patients who relapsed. We denote the pair-wise sampling probability for any two subjects (i, j, with $i \neq j$) by π_{ij} (with $i, j = 1 \ldots m$), conditional on being in the first phase. The sampling probabilities π_i and π_{ij} must be non-zero for all i, j in the phase I sample and be known for i, j in the phase II sample.

Variance estimates of quantities in (1) have been derived by Särndal and Swensson:[14]

$$\widehat{var}\left[d\hat{N}.(t)\right] = \sum_{i=1}^{m} \sum_{j=1}^{m} \left[\frac{dN_i(t)dN \cdot dN_j(t)}{\pi_i \cdot \pi_j} - \frac{dN_i(t)dN \cdot dN_j(t)}{\pi_{ij}}\right] \tag{2}$$

and

$$\widehat{var}\Big[\hat{Y}.(t)\Big] = \sum_{i=1}^{m}\sum_{j=1}^{m}\left[\frac{Y_i(t)\cdot Y_j(t)}{\pi_i\cdot\pi_j} - \frac{Y_i(t)\cdot Y_j(t)}{\pi_{ij}}\right] \tag{3}$$

These estimates are valid conditional on phase I sample and also under more general sampling designs, where phase I is not a random sample from the population, provided that the sampling probabilities are properly generalized.[14]

In our case, we denote with M_s the number of subjects in each stratum and with $m_s < M_s$ the number of subjects sampled at random in phase II without replacement from the sth stratum. The inclusion probabilities, given the phase I sample, will become $\pi_i = m_s/M_s$ and $\pi_{ij} = \frac{m_s(m_s-1)}{M_s(M_s-1)}, \forall i, j \in s$ with $i \neq j$. Since a stratified sample is a set of simple random samples from each stratum, each with constant π_i and π_{ij}, and m is equal to $m_1 + \cdots + m_s + \cdots + m_S$, the estimators for the total number of events and at risk subjects is just the sum of the estimated totals in each stratum. With a straightforward extension of notation, (1) becomes

$$d\hat{N}.(t) = \sum_{s=1}^{S}\sum_{i=1}^{m_s} dN_{is}(t)/\pi_i \quad \text{and} \quad \hat{Y}.(t) = \sum_{s=1}^{S}\sum_{i=1}^{m_s} Y_{is}(t)/\pi_i \tag{4}$$

The same applies for the variance estimates (2) and (3).

The KM estimator of survival can then be derived as

$$\hat{S}(t) = \prod_{u\leq t}\Big[1 - d\hat{\Lambda}(u)\Big] = \prod_{u\leq t}\left[1 - \frac{d\hat{N}.(u)}{\hat{Y}.(u)}\right] \tag{5}$$

where $\Lambda(u)$ represents the cumulative hazard function.

As far as the variance estimate, we concentrate on the variance for the cumulative hazard and, given that $S(t) = exp[-\Lambda(t)]$, the variance of the survival estimate will be then estimated, as usual, according to the delta method by $\widehat{var}[\hat{S}(t)] = \hat{S}(t)^2 \cdot \widehat{var}[\hat{\Lambda}(t)]$. A naïve variance estimator would be the martingale-based variance estimate of $\hat{\Lambda}(t)$,[15] where number of events and subjects at risk are expressed as in (4).

This however represents the irreducible minimum uncertainty that would remain if everyone in the cohort would be sampled in phase II, which is called the contribute of phase I sampling, indicated by $var_I[\hat{\Lambda}(t)]$

$$var_I[\hat{\Lambda}(t)] = \int_0^t \frac{d\hat{N}.(u)}{\hat{Y}.(u)^2} \tag{6}$$

To obtain an unbiased variance estimate, the contribute of phase II sampling has to be added in order to account for the fact that the additional covariate has been ascertained only in the phase II sample. Thus, the variance estimate in two-phase studies results as the sum of these two components

$$var[\hat{\Lambda}(t)] = var_I[\hat{\Lambda}(t)] + var_{II}[\hat{\Lambda}(t)] \tag{7}$$

The phase II estimate makes use of the results of Särndal expressed in (2) and (3). In particular, the contribute of phase II can be estimated as

$$\widehat{var}_{II}\left[\hat{\Lambda}(t)\right] = \sum_{u,u'<t} \widehat{cov}[d\hat{\Lambda}(u), d\hat{\Lambda}(u')] \tag{8}$$

that can be further decomposed into the sum of covariances between total number of subjects at risk and events

$$\begin{aligned}\widehat{cov}[d\hat{\Lambda}(u), d\hat{\Lambda}(u')] = {} & \frac{\widehat{cov}[d\hat{N}.(u), d\hat{N}.(u')]}{\hat{Y}.(u)\hat{Y}.(u')} + \frac{d\hat{N}.(u)d\hat{N}.(u')\widehat{cov}[\hat{Y}.(u), \hat{Y}.(u')]}{\hat{Y}.(u)^2\hat{Y}.(u')^2} + \\ & - \frac{d\hat{N}.(u')\widehat{cov}[d\hat{N}.(u), \hat{Y}.(u')]}{\hat{Y}.(u)\hat{Y}.(u')^2} - \frac{d\hat{N}.(u)\widehat{cov}[d\hat{N}.(u'), \hat{Y}.(u)]}{\hat{Y}.(u')\hat{Y}.(u)^2}\end{aligned} \tag{9}$$

In supplementary material A, we formally show the derivation of the proposed variance (8) using the functional delta method. An alternative derivation can be performed by following the linearization approach by Demnati and Rao[16] (also denoted influence function approach), as shown in supplementary material B. This yields the same estimate, but this latter approach is computationally advantageous, as it does not require the estimation of all the covariance terms in (9). In supplementary material C, we show that variance components are identical to those obtained by a pseudo-likelihood approach on $d\Lambda(t)$.

2.3 Model-based approach

A Cox model adapted for two-phase studies has been developed.[17] To briefly summarize, estimation and testing are based on weighted partial likelihood score equations. The model specifies that the hazard function of the failure time T associated with a vector of possibly time-varying covariates X satisfies

$$d\Lambda(t; X, \beta) = d\Lambda_0(t)exp(\beta X) \tag{10}$$

where $d\Lambda_0(t)$ is the unspecified baseline hazard function and β is the vector of regression parameters. The estimate of the baseline survival curve is $d\hat{\Lambda}_0(t) = \frac{\sum_{s=1}^{S}\sum_{i=1}^{m_s} dN_{is}(t)/\pi_i}{\sum_{s=1}^{S}\sum_{i=1}^{m_s} e^{x_{is}\hat{\beta}} Y_{is}(t)/\pi_i}$ and its variance, derived following the same approach adopted in (8), is implemented in the survey package (with in an unpublished note by Lumley). The test on the regression coefficient β allows to compare survival curves/hazards in different groups properly accounting for the two-phase design.

2.4 Software

The code to compute the KM estimator in a two-phase study has been developed in R software (http://cran.r-project.org/) and is available in the survey package by svykm function,[18] where the variance estimate is performed by using the functional delta method. This can be very slow for large datasets, while the linearization method yields the same estimate in less time. Code is reported in the supplementary material D.

Of note, in the survey package, the phase I variance ($var_I[\Lambda(t)]$) is performed by using the same approach as in (8), with sampling probabilities set to 1 for all subjects (equal probability sampling) and yet it could also be estimated by standard survival approaches giving very similar results.

In the same package, the function svycoxph allows to fit the Cox model adapted to two-phase studies.

3 Simulations

3.1 Simulations protocol

In order to evaluate the performance of the proposed estimator in two-phase studies, we ran a simulation under different sampling schemes. We considered a constant hazard rate of 0.1, yielding the survival function $S(t) = e^{-0.1t}$. Under this setting, we expected a survival of 82% at $t = 2$. In order to allow for the presence of censoring, censoring time was generated according to uniform distribution on (0.5,30.5) and (0.5,10.5) leading to around 5% and 15% of censoring proportion before $t = 2$, respectively.

We drew $B = 1000$ random first-phase samples of size $M = 1000$, from which we sampled a second-phase sample ($m = 50$ and 100 units) according to different sampling schemes in order to mimic the following different study designs:

(1) random sample;
(2) case–control sampling: we randomly sampled $m/2$ individuals among those who experienced the event (cases) up to time 2 and $m/2$ individuals among the others (controls);
(3) stratified sampling: we considered the phase I sample divided into four strata defined by the variable $X = \{0,1\}$ (with frequencies 80% and 20%, respectively) and by the occurrence of the event. The marginal hazard rates in the two strata with $X = 0$ and with $X = 1$ were assumed to be 0.16 and 0.4, respectively. An equal number of subjects ($m/4$) was sampled for each strata (balanced sampling).

$B = 1000$ was chosen in order to get a 5% level of accuracy in the estimate of survival ($S(t)$, $t > 0.3$) in about 95% of the samples. For each sample, $\hat{S}(t)$ has been computed by (5) (using the survey package), and the mean of the estimates over the B simulations ($\bar{\hat{S}}(t)$) has been compared with $S(t)$ in order to assess bias:

- absolute bias $= \bar{\hat{S}}(t) - S(t)$,
- relative bias $= \frac{absolute\,bias}{S(t)}$,
- standardized bias $= \frac{absolute\,bias}{SE(\hat{S}(t))}$,

where $SE(\hat{S}(t))$ is the empirical standard error of the estimate over the B simulations. The mean square error (MSE) was also derived as $(absolute\,bias)^2 + [SE(\hat{S}(t))]^2$. In order to evaluate the variance estimate, for each sample we computed the standard error of $\hat{S}(t)$ according to (7), finding the average very close to the empirical standard error (see Figure 1), as expected. For each simulation, we also computed the 95% confidence interval (CI) of $\hat{S}(t)$ on the logarithm scale to evaluate coverage and length.

In order to show the importance of properly considering the design, we also estimated $S(t)$ and its variance with two naïve approaches. In the first one, we completely ignored the sampling

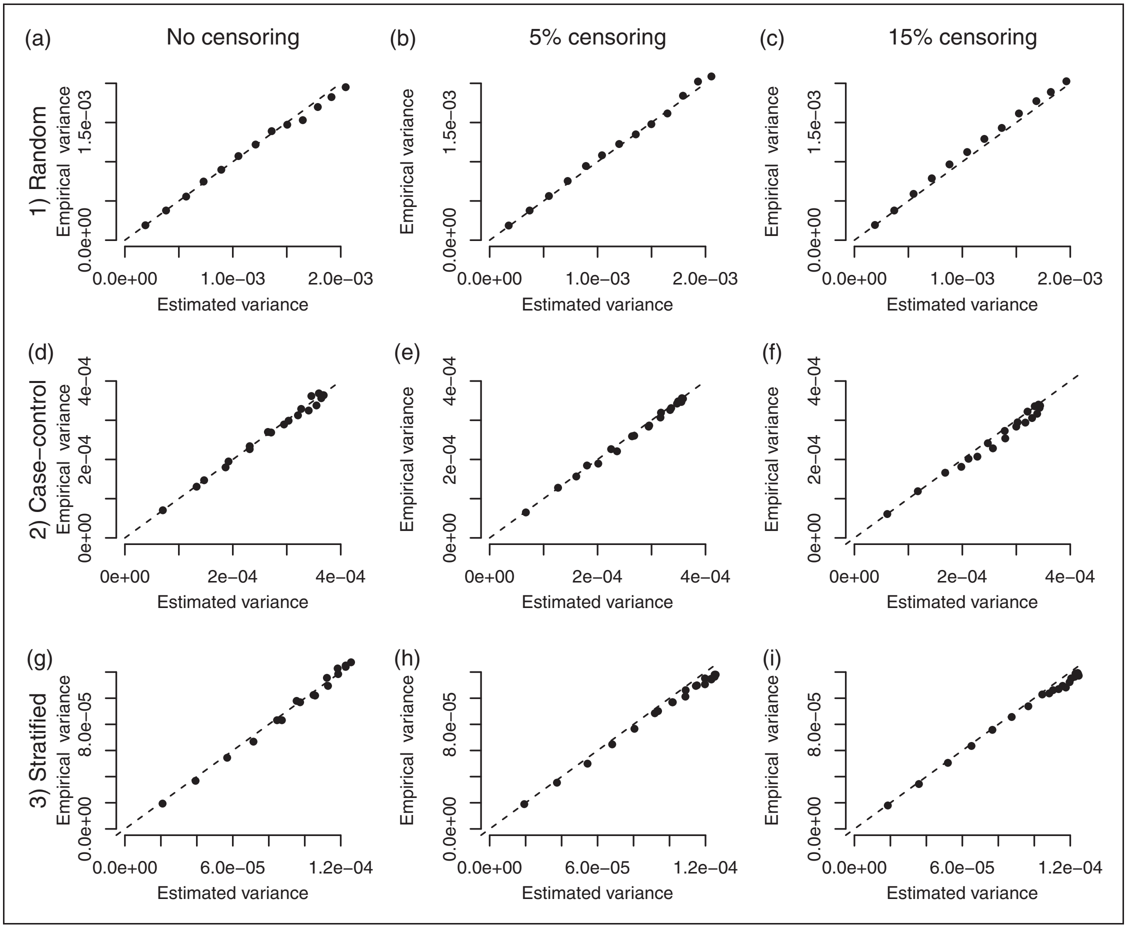

Figure 1. Comparison between estimated and empirical variance under random (panels a, b, and c), case–control (panels d, e, and f), and stratified sampling (panels g, h, and i). Sample size of 50 was considered. Column refers to the amount of censoring: without censoring (first column) and with 5% and 10% censoring (second and third column, respectively). Dashed lines represent the main bisector corresponding to equality between estimated and empirical reference.

scheme/design and used the usual estimators

$$\hat{S}(t) = \prod_{u \leq t}\left[1 - \frac{\sum_{i=1}^{m} dN_i(u)}{\sum_{i=1}^{m} Y_i(u)}\right] \quad \text{and} \quad \widehat{var}(\hat{S}(t)) = \hat{S}(t)^2 \int_0^t \frac{\sum_{i=1}^{m} dN_i(u)}{\left[\sum_{i=1}^{m} Y_i(u)\right]^2}$$

In the second one, we properly estimated $S(t)$ by (5), but we considered only the contribute of phase I sampling by (6) for the variance estimate.

3.2 Simulations results

At first, we evaluated the performance of the proposed estimator in the random sampling scenario (point 1 above). As expected, survival estimates were equal when calculated according to svykm (sampling weights set to 1) or to the standard KM estimator (by the survfit function in the survival package). Comparing the delta methods estimator (8) with Greenwood and martingale-based estimators of variance in (7) gave results that differed very slightly, at most 0.0004. In particular, the delta-based variance was slightly lower and it was closer to the martingale-based variance estimate than to the Greenwood estimate (data not shown).

Results of simulations are reported in Figures 2 and 3 for case–control (point 2) and stratified sampling (point 3), respectively (phase II sample size $m = 50$). The upper panels of the figures report the absolute bias that fluctuates around 0 in each scenario and it is always lower than 0.2%, and the MSE, that increases with time.

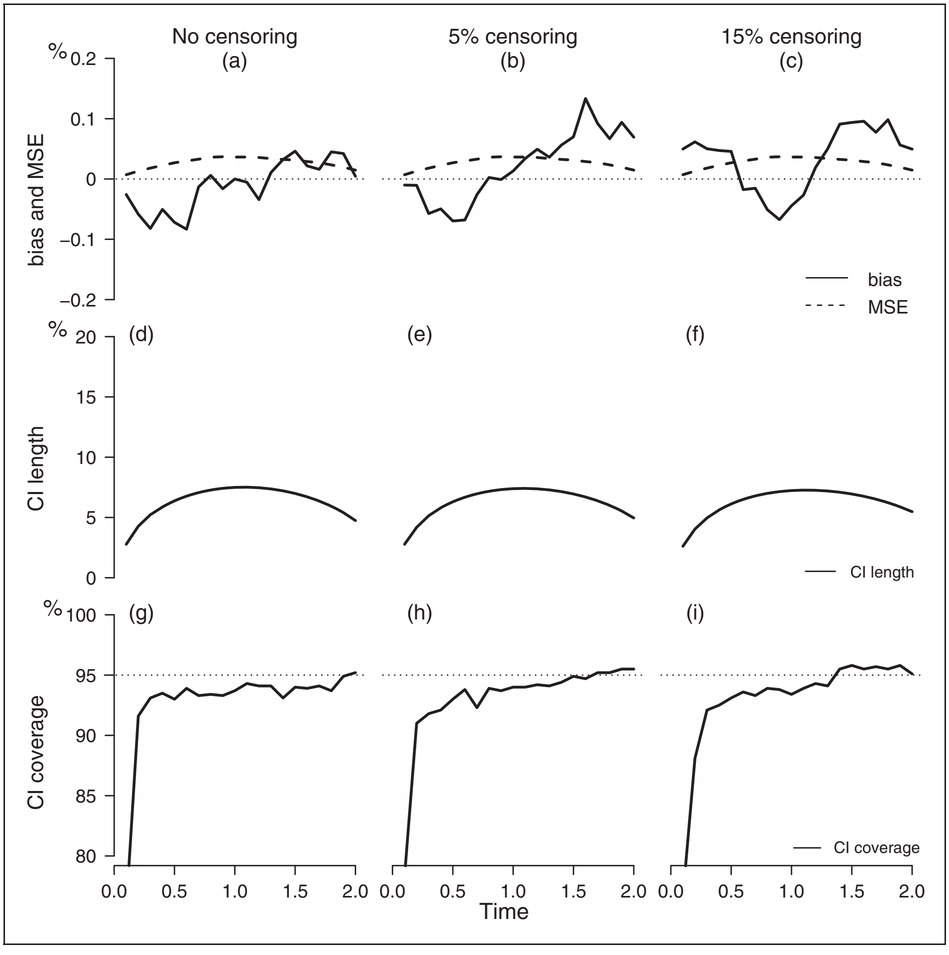

Figure 2. Simulation results for case–control sampling ($m = 50$) without censoring (first column) and with 5% and 10% censoring (second and third column, respectively). Upper graphs report the absolute bias with a solid line and the mean square error (MSE) with a dashed line (panels a, b, and c), the second row reports the confidence interval (CI) length (panels d,e,f), and the third row the CI coverage (panels g, h, and i). Dotted lines represent the reference for bias (bias equal to 0, in panels a, b, and c) and for CI coverage (nominal 95%, in panels g, h, and i).

The relative and the standardized biases are not shown, but resulted to be always lower than 5%. The average length of the CI was higher for random sampling (as expected, data not shown) and was the lowest in the stratified design underlying the advantages of a careful design (Figures 2 and 3, middle panels). The coverage was very close to the nominal 95% value, ranging mostly within a minimum of 92% and a maximum of 96%, except for very early times (bottom panels of Figures 2 and 3).

Under the same settings, we also considered a longer follow-up and confirmed the performance of our survival and variance estimators, with similar results for the different sampling schemes (data not shown).

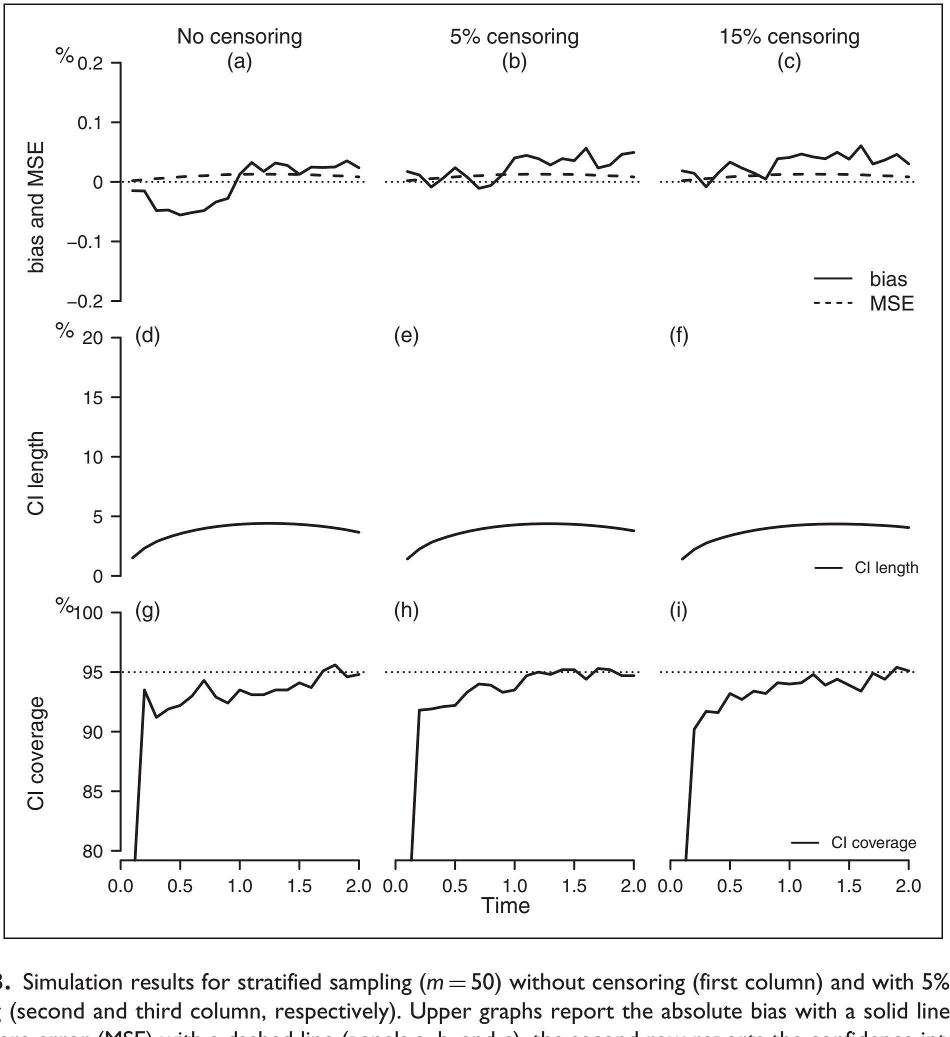

Figure 3. Simulation results for stratified sampling ($m = 50$) without censoring (first column) and with 5% and 10% censoring (second and third column, respectively). Upper graphs report the absolute bias with a solid line and the mean square error (MSE) with a dashed line (panels a, b, and c), the second row reports the confidence interval (CI) length (panels d, e and f), and the third row the CI coverage (panels g, h, and i). Dotted lines represent the reference for bias (bias equal to 0, in panels a, b, and c) and for CI coverage (nominal 95%, in panels g, h, and i).

As far as the naïve approaches, by completely ignoring the design, the absolute bias was huge, as expected. It grew in time starting from around 2% at the beginning of the follow-up and approaching 30% at 2 units of time. For example, in the case–control setting (at point 2 of 3.1), with 15% of censoring, we obtained an estimate of $\hat{S}(2) = 48\%$ with $S(2) = 82\%$. When properly considering the design for the point estimate, but ignoring the contribute of phase II sampling in the variance estimate, the coverage was as low as 50%.

4 Application to a clinical study

We will first describe the adopted sampling design and subsequently the estimated quantities of interest for the study on the role of the GST-T1 gene as a prognostic marker of relapse in childhood ALL.

The AIEOP-ALL2000 clinical trial, beside providing clinical information on the full cohort of 1999 patients (phase I), allowed also to perform an optimal/efficient design for the choice of the subsample to be genotyped (phase II). Given the study objective, we aimed at maximizing the precision of the estimate of the association between genotype and relapse-free interval. For this reason, we applied the optimal sampling strategy developed by Reilly[13] by using the dedicated software.[19] As first, the whole cohort of 1999 patients was divided in six strata according to the event of interest (relapse/no relapse) and to the risk/treatment group (standard, medium, and high risk), as shown in Table 1. Different sampling fractions were applied to each strata, and these were chosen to minimize the variance of the estimate of the association between GST-T1 and relapse in a pilot study of 164 genotyped patients. As typically obtained with this approach,[13] the optimality is achieved by sampling all cases (relapses) and a variable proportion of controls (non relapses) depending on the genetic variability within each strata. Specifically, 13.2%, 25.8%, and 63.8% of non relapsed patients in the standard-, medium-, and high-risk groups were obtained, respectively. Out of the 766 children for whom genotyping was required according to study design, 614 had stored DNA material and valid genotype measure was obtained for 601 patients. Table 1 reports the distribution of patients in the six strata, both for phase I and II samples, and the effective sampling fractions in each stratum.

Because of the design, the subcohort was not representative of the entire cohort and thus sampling weights were adopted in order to recover the representativeness of the subsample, as described in section 2. In practice, each genotyped individual is considered to represent the similar nongenotyped individuals in the same stratum of the whole cohort. This is done by using

Table 1. Distribution of phase I (M_s) and II (m_s) samples in the six strata and sampling fractions expressed as percentages in parenthesis for phase II, ALL patients diagnosed in Italian centers between September 2000 and July 2006.

	Risk group stratification			
	Standard	Medium	High	
		$m_s/M_s(\%)$		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)	245/360
Total	75/515	340/1173	186/311	601/1999

weights equal to the inverse of the probability of being sampled from the cohort. For example, the 54 genotyped children in the standard-risk group (Table 1) who did not relapse are weighted by the inverse of the sampling fraction (i.e., by 1/0.111), so that they contribute to statistical analysis for the corresponding 487 patients in the full cohort.

The performance of the proposed estimate can be actually shown here by deriving the overall relapse-free survival for patients in phase II sample, as well as for the whole cohort of 1999 patients. Our phase II estimate was found to be very similar to the cohort estimate: three-year relapse-free interval of 86.7% (standard error (SE) 0.84%) and 86.2% (SE 0.83%), respectively. Moreover, the efficiency of the design can be expressed through the ratio between the phase I variance and the total variance of the KM estimate, which accounts for the fact that only a subsample was genotyped. In our study, this ratio was about 91% ((0.00638/0.00699)% at three years), which is very satisfactory given that the results is obtained after genotyping only 30% of the cohort.

We report here the R code used to compute this estimate:

```
library(survey)
d.std < − twophase(id = list(∼ upn, ∼ upn), subset =∼ !is.na(GST_T),
strata = list(NULL, ∼ interaction(rel, elfin)), data = dat)
GSTse < − svykm(Surv(timerel, evento == 1) ∼ 1, d.std, se = TRUE)
```

The twophase function describes the design and produces a survey object, while the svykm function performs the estimate by the delta approach. Details on the use of these functions can be found in Lumley.[9,18]

Relapse-free survival by GST-T1 deletion can only be estimated based on phase II data by the code:

```
GST1 < − svykm(Surv(timerel, evento == 1) ∼ GST_T, d.std, se = TRUE)
```

and shows that patients with GST-T1 deletion had higher risk of relapse, with five-year relapse-free interval 75.7% (SE 4.1%) versus 83.5% (SE 1.1%, Figure 4). The computational time to estimate these standard errors was 21 min. The same result was obtained in 15 min by using the new code in supplementary material D:

```
GST1l < − svykm2(Surv(timerel, evento == 1) ∼ GST_T, d.std, se = TRUE)
```

Of note, we also applied the usual KM estimator ignoring the design and, as expected, found very low relapse-free intervals (44.1% for deleted and 55.8% for nondeleted subjects; see gray lines in Figure 4). When we naïvely estimated variance considering only the contribute of phase I sampling, we obtained lower standard errors: at five-year SE for deleted patients dropped to 2.9%, while for not deleted was 0.9%.

The Cox model adapted for the two-phase design and applied to the cause specific hazard of relapse:[17]

```
cox < − svycoxph(Surv(timerel, evento == 1) ∼ GST_T, d.std)
```

allows to test the null hypothesis $\beta = 0$ for genotype (under the assumption of proportional hazard). The Wald test statistics is $z = 1.969$, thus resulting in a borderline significant difference at 5% level

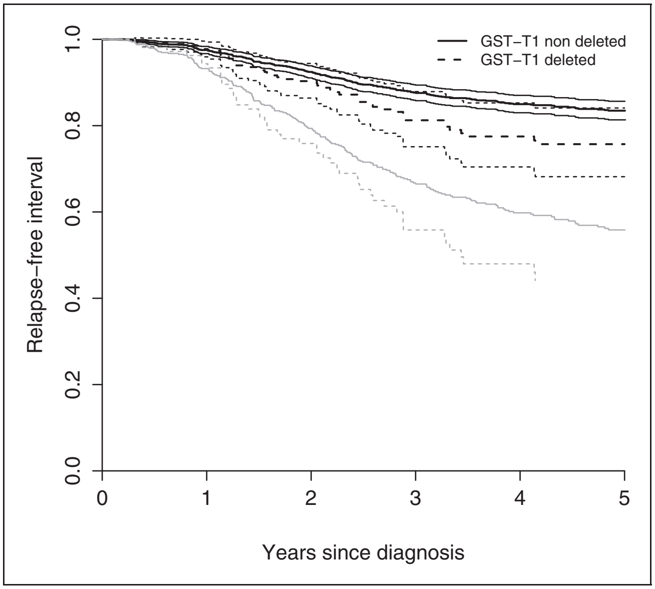

Figure 4. Estimated relapse-free interval and point-wise confidence intervals in patients with normal (reported with a solid line) and deleted (dashed line) GST-T1 gene. Gray color represents estimates obtained ignoring the design.

(p-value 0.049). Survival curves estimated by this model agreed well with the curves in Figure 4 (data not shown).

After adjusting for relevant factors, the Cox model gives an hazard ratio of 1.32 (95% CI: 0.90, 2.00, p-value = 0.14) for GST-T1 deleted patients versus nondeleted.[12]

5 Discussion

Two-phase studies are attractive for their economy and efficiency, especially in research settings where many novel biomarkers need to be investigated. Efficiency is achieved by optimally drawing from a cohort a possibly biased subsample, where the biomarker is measured, and by applying appropriate inference methods for two-phase designs, that make use also of the cohort information in order to obtain unbiased estimates. These designs have been applied sporadically mainly because there is no consolidated experience both in optimal sampling and statistical analysis and there is limited availability of specific statistical software. The two-phase design is particularly advantageous, compared to matched case–control studies nested in a cohort, since it allows to estimate the impact of a biomarker on event incidence and also to consider different types of events.

In this work, we extended the estimator of survival probability and its variance to deal with two-phase designs and showed that they have a good performance under different simulated scenarios.

The code to compute these estimates has been developed in R software and is available in the survey package by svykm function.[18] This function uses the functional delta method that can be very slow for large dataset, and we provided new code applying the linearization method that yields the same estimate in less time. The linearization strategy is similar to the one adopted by Williams to derive a variance estimator for the product-limit estimator of survival under a cluster sampling design.[20] It is also related to the pseudo-likelihood approach by Samuelsen.[1]

In our motivating context, two-phase design is particularly attractive for a parsimonious use of biospecimens collected at diagnosis and also for the potential to study the association of GST-T1 deletion with different outcomes. The main objective was to study the association with the risk of relapse or the complementary relapse-free interval. In the design stage, we proposed an optimal approach for the choice of sampling fractions that is based on relapse as the event of interest. This choice was actually quite efficient, as seen in the comparison between phase I and II estimates and from the design effect on the variance estimate (section 4). Differently from nested case–control studies, we might be able to analyze the incidence of a different event, for instance the combined endpoint (relapse or toxicity) or just toxicity. However, the efficiency could be lower unless the subsampling is adapted to this new endpoint by including a further strata on toxicity in the sampling process.

In order to recover the representativeness of the subcohort for the entire cohort, we used weights related to the inverse of the probability to be sampled, but more general weights can be used in this approach, such as calibration weights.[9,21] 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 motivating context, it is unlikely to find any strong relation between phase I and II variables; therefore, no big advantage would be expected by calibration.

The proposed estimator has been developed under a two-phase design, but it can also be applied in different settings. The most obvious example is a (not-matched) case–control study, where the phase I sample has two strata: events (cases) and not events (controls); in the phase II, all the cases and a random sample of controls are sampled.[22,23] Furthermore, it can be applied under a general biased sample, where the extent and direction of bias is known, or in the presence of missing data, where the variables affecting the missingness can be reasonably identified.[9,14]

The proposed estimator agrees well with the estimators proposed by Mark and Katki[24] for cohorts data with missing covariate information, in which sampling/missingness probabilities are modeled through a generalized linear model.[24]

In this work, a valid survival estimator has been proposed for general two-phase sampling designs. Further work will concern developing an incidence estimator that account for possible competing risks.

Acknowledgements

This research greatly benefited by discussion with Thomas Lumley.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the European Commission (FP7-HEALTH-F2-2011 261474).

References

1. Samuelsen SO. A pseudolikelihood approach to analysis of nested case–control studies. *Biometrika* 1997; **84**: 379–394.
2. Salim A, Hultman C, Sparén P, et al. Combining data from 2 nested case–control studies of overlapping cohorts to improve efficiency. *Biostatistics* 2009; **10**: 70–79.
3. Prentice RL. A case–cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**: 1–11.
4. Borgan Ø and Samuelsen SO. A review of cohort sampling designs for Cox's regression model: potentials in epidemiology. *Norsk Epidemiologi* 2003; **13**: 239–248.
5. Langholz B and Borgan ØR. Counter-matching: a stratified nested case–control sampling method. *Biometrika* 1995; **82**: 69–79.
6. Neyman J. Contribution to the theory of sampling human populations. *J Am Stat Assoc* 1938; **33**: 101–116.
7. Breslow NE, Lumley T, Ballantyne CM, et al. Using the whole cohort in the analysis of case–cohort data. *Am J Epidemiol* 2009; **169**: 1398–1405.
8. Breslow NE and Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *J Royal Stat Soc C* 1999; **48**: 457–468.
9. Lumley T. *Complex surveys: A guide to analysis using R.* Hoboken, NJ: John Wiley & Sons, 2010, p.296.
10. Haneuse S, Saegusa T and Lumley T. osDesign: an R package for the analysis, evaluation, and design of two-phase and case–control studies. *J Stat Software* 2011; **43**: 1–29.
11. Stanulla M, Schrappe M, Brechlin AM, et al. Polymorphisms within glutathione S-transferase genes (GSTM1, GSTT1, GSTP1) and risk of relapse in childhood B-cell precursor acute lymphoblastic leukemia: a case–control study. *Blood* 2000; **95**: 1222–1228.
12. Franca R, Rebora P, Basso G, 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**: 1905–1916.
13. Reilly M. Optimal sampling strategies for two-stage studies. *Am J Epidemiol* 1996; **143**: 92–100.
14. Särndal CE and Swensson B. A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Int Statl Rev* 1987; **55**: 279–294.
15. Tsiatis AA. A large sample study of Cox's regression model. *Ann Stat* 1981; **9**: 93–108.
16. Demnati A and Rao JNK. Linearization variance estimators for model parameters from complex survey data. *Survey Methodol* 2010; **36**: 193–201.
17. Lin DY. On fitting Cox's proportional hazards models to survey data. *Biometrika* 2000; **87**: 37–47.
18. Lumley T. Analysis of complex survey samples. *J Stat Software* 2004; **9**: 1–19.
19. Reilly M, Salim A and Mahmud S. Software tools for optimal two-stage sampling a user guide, http://www.meb.ki.se/marrei/software/userguide.pdf (2000, accessed 29 November 2012).
20. Williams RL. Product-limit survival functions with correlated survival times. *Lifetime Data Anal* 1995; **1**: 171–186.
21. Scott AJ and Wild CJ. Fitting regression models with response-biased samples. *Can J Stat* 2011; **39**: 519–536.
22. Breslow NE. Statistics in epidemiology: the case–control study. *J Am Stat Assoc* 1996; **91**: 14–28.
23. Langholz B. Case–control studies = odds ratios: blame the retrospective model. *Epidemiology* 2010; **21**: 10–12.
24. Mark SD and Katki HA. Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (nested) cohort studies with missing case data. *J Am Stat Assoc* 2006; **101**: 460–471.