

Optimal two-phase design and incidence estimation in cohort studies

Disegno ottimale a due fasi e stima dell'incidenza di evento in studi di coorte

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. In order to estimate event incidence in the subcohort, a Kaplan-Meier method accounting for the design has been derived with a proper variance estimator. The proposed method is applied in the context of a two-phase study on childhood acute lymphoblastic leukaemia.

Abstract I disegni di studio a due fasi sono attraenti per la loro efficienza quando si hanno a disposizione ampie coorti per valutare il ruolo predittivo e prognostico di fattori genetici e biologici rari. In questo tipo di studi, l'informazione genetica/biologica è raccolta solo su un campione ottimale (fase II) della coorte (fase I). Per stimare l'incidenza di evento nella sottocoorte, lo stimatore di Kaplan-Meier (e la sua varianza) sono stati adattati per tener conto del disegno di studio. Il metodo proposto è stato applicato per stimare l'incidenza di ricaduta in uno studio a due fasi sulla leucemia linfoblastica acuta pediatrica.

Key words: two-phase design, survival, optimal sampling, case-control, missing data

Paola Rebora

Department of Health Sciences; University of Milano-Bicocca; Monza, Italy, e-mail: paola.rebora@unimib.it

Maria Grazia Valsecchi

Department of Health Sciences; University of Milano-Bicocca; Monza, Italy e-mail: grazia.valsecchi@unimib.it

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, that although is very efficient, allows to evaluate risk/protective factors for the event type that defines the cases (and not for other possible events of interest). This considerably limits the potentiality of the study, in particular when precious samples from bio-banks are used to measure additional 'costly' covariates.

An alternative approach is to use a two-phase sampling technique [1]. 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. The aim of this work is to develop a weighted Kaplan-Meier (KM) survival estimator and its variance in the framework of a two-phase design.

2 Methods

Let T_i be the failure time and C_i the censoring time of subject i ($i = 1 \dots M$) 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. $dN_i(t) = I(Z_i = t)\Delta_i$) and $Y_i(t)$ indicates whether the subject is at risk and under observation at time t ($Y_i(t) = I(Z_i \geq t)$).

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 $S(t) = P(T > t)$ in subgroups defined according to the variable ascertained only in the phase II sample. Under efficient sampling adopted in phase II the Kaplan-Meier estimator would obtain a biased estimate. 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 $d\hat{N}_\cdot(t) = \sum_{i=1}^m dN_i(t)/\pi_i$ and $\hat{Y}_\cdot(t) = \sum_{i=1}^m Y_i(t)/\pi_i$, 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). Their variance estimates have been derived by Särndal [6].

The Kaplan-Meier estimator of survival can then be derived as:

$\widehat{S}(t) = \prod_{u \leq t} [1 - d\widehat{\Lambda}(u)] = \prod_{u \leq t} [1 - \frac{d\widehat{N}(u)}{\widehat{Y}(u)}]$ 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 according to the delta method by: $\widehat{var}[\widehat{S}(t)] = \widehat{S}(t)^2 \widehat{var}[\widehat{\Lambda}(t)]$.

A naïve variance estimator would be the martingale-based variance estimate of $\widehat{\Lambda}(t)$ [7] where number of events and subjects at risk are properly weighted. 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: $var_I[\widehat{\Lambda}(t)] = \int_0^t \frac{d\widehat{N}(u)}{\widehat{Y}(u)^2}$. 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: $var[\widehat{\Lambda}(t)] = var_I[\widehat{\Lambda}(t)] + var_{II}[\widehat{\Lambda}(t)]$. The phase II estimate makes use of the results of Särndal and can be estimated as: $\widehat{var}_{II}[\widehat{\Lambda}(t)] = \sum_{u, u' < t} \widehat{cov}[d\widehat{\Lambda}(u), d\widehat{\Lambda}(u')]$, that can be further decomposed into the sum of covariances between total number of subjects at risk and events:

$$\begin{aligned} \widehat{cov}[d\widehat{\Lambda}(u), d\widehat{\Lambda}(u')] = & \frac{\widehat{cov}[d\widehat{N}(u), d\widehat{N}(u')]}{\widehat{Y}(u)\widehat{Y}(u')} + \frac{d\widehat{N}(u)d\widehat{N}(u')\widehat{cov}[\widehat{Y}(u), \widehat{Y}(u')]}{\widehat{Y}(u)^2\widehat{Y}(u')^2} + \\ & - \frac{d\widehat{N}(u')\widehat{cov}[d\widehat{N}(u), \widehat{Y}(u')]}{\widehat{Y}(u)\widehat{Y}(u')^2} - \frac{d\widehat{N}(u)\widehat{cov}[d\widehat{N}(u'), \widehat{Y}(u)]}{\widehat{Y}(u')\widehat{Y}(u)^2} \end{aligned} \quad (1)$$

The formal derivation of the proposed variance used the functional delta method. An alternative derivation can be performed by following the linearisation approach by Demnati and Rao [2] (also denoted influence function approach). 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 (1). An identical estimator would be also obtained by a pseudo-likelihood approach.

3 Simulations

In order to evaluate the performance of the proposed estimator in two-phase studies, we ran a simulation study, where we considered a constant hazard rate of 0.1, and censoring time was generated according to uniform distribution. 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: i) random sample; ii) 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); iii) strat-

ified sampling: we considered the phase I sample divided into 4 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).

Bias fluctuated around 0 in each scenario and was always lower than 0.2%. For each simulation, the 95% confidence interval of $\hat{S}(t)$ was computed on the logarithm scale ($\log(S(t)) = -\Lambda(t)$) to evaluate coverage and length. The average length of the confidence interval was higher for random sampling (as expected) and was the lowest in the stratified design underlying the advantages of a careful design. Specifically, with a second phase sample of 50 subjects, the average length of confidence interval was 0.15, 0.06 and 0.04 for random (i), case-control (ii) and stratified samples (iii), respectively. The coverage was very close to the nominal 95% value, ranging mostly within a minimum of 92% and a maximum of 96%.

4 Application to a clinical study

The clinical context that motivated this work is represented by a study on childhood acute lymphoblastic leukemia, performed in order to evaluate the role of different genetic polymorphism on treatment failure due to relapse. We will concentrate on the homozygous deletion (or null genotype) in glutathione S-transferase- θ (GST-T1) gene. 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 [3]. 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 [5]. 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, high risk). 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 [5], 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. 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.

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

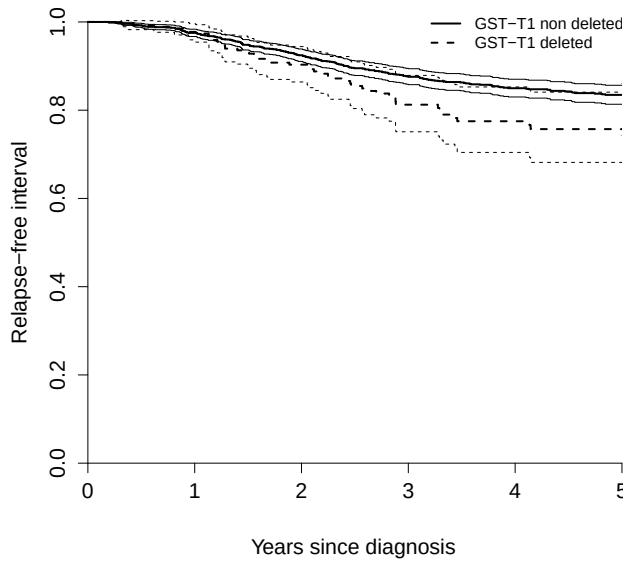


Fig. 1 Estimated relapse-free interval and point-wise confidence intervals in patients with normal (reported with a solid line) and deleted (dashed line) GST-T1 gene.

similar to the cohort estimate: 3-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%, which is very satisfactory given that this result is obtained after genotyping only 30% of the cohort.

Relapse-free survival by GST-T1 deletion can only be estimated based on phase II data and shows that patients with GST-T1 deletion had higher risk of relapse, with 5-year relapse-free interval 75.7% (SE 4.1%) versus 83.5% (SE 1.1%, Figure 1). A Cox model adapted for two-phase design, when applied to the cause specific hazard of relapse, gives an hazard ratio of 1.32 (95% CI: 0.90, 2.00) for GST-T1 deleted patients versus non deleted, after adjusting for relevant factors [3].

5 Discussion

Two-phase studies are attractive for their economy and efficiency, especially in research settings where many novel biomarkers need to be investigated. 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. 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 this estimates has been developed in R software and is available in the `survey` package by `svykm` function [4], but can be very slow for large data-set. We also developed a new code applying the linearisation method, that yields the same estimate in less time.

In our motivating context, two-phase design is particularly attractive for a parsimonious use of bio-specimens collected at diagnosis and also for the potential to study the association of GST-T1 deletion with different outcomes. 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.

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 [4]. When phase II variables are common genetic polymorphism, 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.

In this work, a valid survival estimator has been proposed for general two-phase sampling designs. We are working to develop an incidence estimator that account for possible competing risks.

References

1. Breslow, N., Lumley, T., Ballantyne, C., Chambless, L., Kulich, M.: Using the Whole Cohort in the Analysis of Case-Cohort Data. *Am J Epidemiol* **169**(11), 1398 – 1405 (2009). DOI 10.1093/aje/kwp055
2. Demnati, A., Rao, J.: Linearization variance estimators for model parameters from complex survey data. *Survey Methodology* **36**, 193 – 201 (2010)
3. Franca, R., Rebora, P., Basso, G., Biondi, A., Cazzaniga, G., Crovella, S., Decorti, G., Fagioli, F., Giarin, E., Locatelli, F., 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* **13**(16), 1905–1916 (2012)
4. Lumley, T.: *Complex Surveys: A Guide to Analysis Using R*. Wiley Series in Survey Methodology. John Wiley & Sons, Hoboken, New Jersey (2010)
5. Reilly, M.: Optimal sampling strategies for two-stage studies. *Am J Epidemiol* **143**(1), 92–100 (1996)
6. Särndal, C., Swensson, B.: A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. *International Statistical Review* **55**(3), 279 – 294 (1987)
7. Tsiatis, A.: A large sample study of Cox’s regression model. *Ann Stat* **9**(1), 93 – 108 (1981)