Sampling Strategies to Evaluate the Prognostic Value of a New Biomarker on a Time-to-Event End-Point

Francesca Graziano
University of Milan–Bicocca: Universita degli Studi di Milano-Bicocca
Maria Grazia Valsecchi
University of Milano-Bicocca
Paola Rebora (paola.rebora@unimib.it)
University of Milano-Bicocca https://orcid.org/0000-0003-0606-5852

Research article

Keywords: case-control design, cohort studies, power, two-phase sampling, weighted Cox model

Posted Date: November 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-102892/v1

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on April 30th, 2021. See the published version at https://doi.org/10.1186/s12874-021-01283-0.

TITLE: Sampling strategies to evaluate the prognostic value of a new biomarker on a time-to-event end-point

RUNNING TITLE: Two phase sampling strategies

Francesca Graziano[1], Maria Grazia Valsecchi[1] and Paola Rebora[1*]

Author affiliations: [1] BICOCCA BIOINFORMATICS BIOSTATISTICS AND BIOIMAGING CENTRE-B4, School of Medicine and Surgery, University of Milano - Bicocca, Via Cadore 48, 20900 Monza, Italy

*Correspondence:

paola.rebora@unimib.it, Tel. +39 02 6448 8165

BICOCCA BIOINFORMATICS BIOSTATISTICS AND BIOIMAGING CENTRE-B4, School of Medicine and Surgery, University of Milano - Bicocca, Via Cadore 48, 20900 Monza, Italy

ABSTRACT

Background: The availability of large epidemiological or clinical data storing biological samples allow to study the prognostic value of novel biomarkers, but efficient designs are needed to select a subsample on which to measure them for parsimony and economical reasons. Two-phase stratified sampling is a flexible approach to perform such sub-sampling, but literature on stratification variables to be used in the sampling and power evaluation is lacking especially for survival data.

Methods: We compared the performance of different sampling designs to assess the prognostic value of a new biomarker on a time-to-event end-point, applying a two-phase Cox model weighted by the inverse of the empirical inclusion probability.

Results: Our simulation results suggest that case-control stratified (or post stratified) by a surrogate variable of the marker can yield higher performances than simple random, probability proportional to size and case-control sampling. In the presence of high censoring rate, results showed an advantage of nested case-control and counter-matching designs in term of design effect, but their constrain in using a fixed ratio between cases and controls might be disadvantageous. On real data on childhood acute lymphoblastic leukemia, we also found that optimal sampling using pilot data is greatly efficient.

Conclusions: Our study suggests that, in our sample, case-control stratified by surrogate and nested case-control yield estimates and power comparable to estimates obtained in the full cohort while strongly decreasing the number of patients required. We recommend to plan the sample size and using sampling designs for exploration of novel biomarker in clinical cohort data.

Keywords: case-control design, cohort studies, power, two-phase sampling, weighted Cox model

1. BACKGROUND

In the past decades, there has been a growing number of epidemiological (1-3) and longitudinal studies storing biological samples (4) to allow retrospective evaluation of new research questions, such as evaluating the prognostic value of new biomarkers. This approach is convenient, significantly reducing the time of the study. However, the analysis of novel biomarkers could be expensive. Sub-sampling strategies result in considerable cost savings and parsimonious use of biological specimens, by restricting data extraction to an informative subgroup of the original sample. Instead of choosing the subgroup at random, it may be carefully sampled to obtain unbiased and more precise results (5, 6).

Two-phase sampling is a general approach to perform such sub-sampling, including case-control and case-cohort designs (7, 8). This approach considers the entire cohort as the first-phase sample from the population of interest. In the second phase, subsamples are drawn from the cohort to measure additional information, such as new biomarkers of interest (9). An optimal sampling strategy was proposed for stratified two-stage studies with binary outcome, however it needs the availability of pilot data on the biomarker of interest that are not always available. Moreover, there is no literature on the choice of variables to be used as strata and no extension for time-to-event outcomes.

The sample size of retrospective studies is often planned considering only budget constrain without a proper evaluation of the statistical power (10), also due to the lack of methodologies to evaluate power in this setting. Cai and Zeng (11) focused on the power in case–cohort design without any stratification; Haneuse et al. (12, 13) focused on binary outcomes, but a general strategy for power evaluation is missing for survival data.

In this study we compared different approaches of sampling designs in the two-phase setting to assess the prognostic value of a new biomarker on a time-to-event end-point and provide a tool to estimate power by simulations. In particular, we focused on the sampling design of the

sub-cohort on which to measure the new biomarker. In this context, we explored the contribution of different types of stratification variables (e.g. surrogate, risk factor, confounder…) and of different sampling designs on power and design effect. We performed a power evaluation varying the sub-cohort sample size and using real data from a study carried out using data from a randomized trial in childhood acute lymphoblastic leukemia (ALL). Briefly, this study was performed to evaluate the role of different genetic polymorphisms on treatment failure due to relapse (14, 15). Clinical data and other information were available for the whole trial cohort and biological samples were stored at diagnosis. The genetic polymorphisms were retrospectively evaluated on these specimens using a two-phase design.

2. METHODS

2.1 Notation settings

Due to a survival analysis framework, commonly survival notation has been used. Let T_i be the failure time and C_i the censoring time of subject i $(i = 1 \dots N)$ in a cohort (phase I) of size N followed-up to time τ. T_i and C_i are assumed to be independent, $T \perp C$. Define $Z = \min(T, C)$ and $\Delta = I(T \leq C)$. Let $h_i(t)$ be the hazard rate for the ith individual. The hazard function, modelled using the Cox proportional hazards model, is equal to $h_i(t) = h_0(t)\exp(\beta X_i)$ where $h_0(t)$ is the baseline hazard, X_i the vector of the explanatory variables for individual i and β's the corresponding regression coefficients. The classical approach for estimating β is to maximize the partial likelihood (16). Suppose that information on one variable of interest, i.e. X_{BM} , is available only for a subset n < N of subjects drawn from the phase I data and let ξ_i indicate whether subject i is selected into this sample. We will refer to the $n = \sum_{i=1}^{N} \xi_i$ subjects as the phase II sample. Let $\pi_i = P(\xi_i = 1 | X_i, \Delta_i, Z_i)$ 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 $(\pi = n/N)$.

However, in a stratified sampling this probability 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).

2.2 Simulation context

First phase sample

To mimic a realistic context, we hypothesized a cohort of subjects of size N (i.e. clinical trial cohort, register, clinical cohort) followed up to τ, in which we aim to evaluate the prognostic value of a new biomarker (X_{BM}) on a time-to-event endpoint (T) in the presence of a possible confounder (X_{Conf}), a risk factor ($X_{Risk\ Fact}$) and a possible auxiliary/surrogate variable (X_{Surr}) of the marker of interest. To describe and illustrate relationships between these, a Directed Acyclic Graph (DAG) was displayed in **Figure 1**. In particular, we assumed the confounder to have an impact on both the biomarker and the event of interest, the risk factor to be associated only with the event of interest, and the surrogate to be associated only with the biomarker. We believe that these variables might well represent the majority of variables usually available in practice, even though in a simplified setting for simulation.

Insert Figure 1

Second Phase Sample

We assumed that the risk factor, the confounder and the surrogate are known for all subjects in the first phase (N), while the biomarker (X_{BM}) is measured only on the subset of n individuals (second phase sample).

To sample the subset a stratified two-phase sampling approach was used. Strata for the first phase sample were defined using the following variables: event, event and risk factor, event and confounder, event and surrogate.

By note, in this work, we consider only sampling done at the end of the follow-up (τ). Subjects who developed the event during the follow-up are defined as cases and subjects event-free at τ as controls.

The sample size of the second phase is fixed (n), but the sampling probabilities depend on different designs, as described below:

(i) **Simple Random Sample (SRS)**: all possible subsamples have an equal probability to be chosen.

(ii) **Probability Proportional to Size (PPS):** The inclusion probability for cases and controls is constant (π_i=n/N) applied within each stratum without replacement (17). Thus, the size for each stratum is given by the total size of the stratum in the original cohort multiplied by n/N.

(iii) **Case-Control (CC)** is performed by separately sampling cases and controls (18). As we aimed to compare different sampling strategies with a fixed sample size, we did not necessarily select all cases from the full cohort as often done. We fixed a total sample size (n) and an equal number of cases (n/2) and controls (n/2) were selected. We also considered stratified CC by using the variables available in phase I (see **Figure 1**): separated simple random sampling was performed in each stratum. A balance design was considered (19).

(iv) **Nested case-control (NCC)** can be considered as a particular case of case-control designs in which controls are sampled by the set of subjects event-free at the time of event of the case (20-22). Sampling probabilities for controls were derived by Samuelsen (23), while for cases they were equal to 1 when the second phase sample size n was at least twice the total number of event in the entire cohort ($\sum_{i=1}^{N} \Delta_i$) and $\pi_i = (n/2)/(\sum_{i=1}^{N} \Delta_i)$ otherwise.

(v) **Counter matching (CM)** is an alternative stratified version of the NCC. In this design, the selection of controls is conducted by sampling from the set at risk in the opposite stratum

at the time of event of the case. Inclusion probabilities for controls within strata were derived by Samuelsen (24) while for cases, the π_i was derived as in NCC design. As the aim is to maximize the "discordance" of exposure within case-controls sets (25-27), the variables used to define strata must be a proxy for the variables of interest, thus we used only the surrogate variable X_{Surr} as strata for this design.

Figure 2 illustrates an example of each sampling design method described above. Specifically, in the upper part of the figure we displayed PPS and CC considering a stratification for a binary variable; in the lower part NCC and CM designs are displayed. By note we considered a 1:1 matching ratio between cases and controls in NCC and CM designs.

Insert Figure 2

2.3 Evaluation of Biomarker impact on the event

The following Cox model was applied to assess the influence of the biomarker on the event adjusting for the confounder variable X_{Conf} (following the minimal set of adjustment suggested in **Figure 1**):

$$h_i(t) = h_0(t) \exp(\beta_{BM} X_{BM_i} + \beta_{Conf} X_{Conf_i}) \quad (1)$$

where β_{BM} and β_{Conf} represent the regression coefficients of the biomarker and of the confounder, respectively. Given the availability of the biomarker only for the sub-cohort (phase II), we applied a two-phase Cox model, in which regression coefficients are estimated by maximizing the partial likelihood weighted by the inverse of the empirical inclusion probability ($w_i = {}^{1}/_{\pi_i}$) that accounts for the specific sampling design (6, 28, 29). In SRS, CC and PPS designs (17, 30) empirical inclusion probabilities (π_i) were calculated using a

standard approach implemented in the twophase function in the `survey package`. Instead, π_i's were calculated following Samuelsen (23) for NCC and following Rivera for CM (25). As surrogate variables are rarely available for new biomarkers at the design stage, we considered also a situation in which the surrogate variable was identified only at the analysis stage. Thus we ran a Cox model post-stratifying for the surrogate variable to take advantage of it in the analysis stage (8).

2.4 Simulations parameters

The performance of the different designs was investigated through simulations. The number of simulated data-set was set at B=2000 assuming a level of accuracy equal to 0.0046 and a variance of the X_{BM} equal to 0.011 with a 5% significance level (31). For each scenario we drew B=2000 random first-phase samples of N=2000 subjects to generate the hypothetical cohort described above. We considered a common and rare marker with frequency in the entire cohort of nearly 25% and 5%, respectively. In order to cover different levels of accuracy of the surrogate in "predicting" the value of the biomarker, we varied the simulation parameter to get different specificity and sensitivity.

The time-to-event end-point was generated (31, 32) from a Weibull hazard model as $T = (-logU/\lambda \exp(\beta' X))^{1/p}$, where p=0.9, λ=0.1, with U following a uniform distribution on the interval from 0 to 1 and with the matrix of covariates X including the biomarker value (X_{BM}), the risk factor ($X_{Risk\ Factor}$), and the confounder (X_{Conf}). The regression coefficients (β) were inspired by the observed values in ALL data (15, 33). We started by simulating the confounder variable as a dichotomous variable with $P(X_{Conf} = 1) = 0.5$; the biomarker was simulated by a binomial distribution with $P(X_{BM} = 1|X_{Conf}) = \exp(a + b * X_{conf}) / (1 + \exp(a + b * X_{conf}))$ resulting in a frequency in the entire cohort of nearly 25% (a=-2 an b=1.7) and 5% (a=-4 and b=1.5) for common and rare biomarker, respectively. The surrogate or

auxiliary variable, using X_{BM} as gold-standard, was simulated as $P(X_{Surr} = 1|X_{BM}) = \exp(c + d * X_{BM}) / (1 + \exp(c + d * X_{BM}))$. In order to cover different levels of accuracy of the surrogate in "predicting" the value of the biomarker, we set different values of parameters c and d. For example, we set c=-0.87 and d=1.74, resulting in specificity, $P(X_{Surr} = 0|X_{BM} = 0)$, and sensitivity, $P(X_{Surr} = 1|X_{BM} = 1)$, values equal to 70%; c= -0.9 and d=2.6 resulting in specificity and sensitivity values equal to 70% and 80%, respectively; c=-1.1 and d=3.3 resulting in specificity and sensitivity values equal to 70% and 90%, respectively. Finally, an additional binary risk factor $X_{Risk\ factor}$ was generated with a probability of $P(X_{Risk\ factor} = 1) = 0.4$. Details of all specific parameters were reported in the Additional file Table S1.

Censoring was generated with an exponential function with different rates ρ = 0, 0.1, 0.4, resulting in 0%, 15% and 50% of censoring at the end of follow-up τ. Minimum between time-to-event T_i and censoring C_i ($Z_i = \min(T_i, C_i)$) was calculated, with $\Delta_i = I(T_i < C_i)$. Administrative censoring was set at τ=2. An average of 500 events for each phase I dataset was reached at the end of follow-up.

To compare the performance of different sampling designs we sampled from each one of the B phase I data, a phase II sample with size n, as described in the chapter "**Second-phase sample**".

Information on X_{BM} was disregarded for subjects not included in the phase II sample and a weighted Cox model was applied to estimate β_{BM} as described in **Evaluation of Biomarker impact on the event** section**.**

The performance of the estimate of β_{BM}over the B simulation has been assessed by the following measures (31):

(i) Absolute bias, given by $= \bar{\hat{\beta}}_{BM} - \beta_{BM}$, where $\bar{\hat{\beta}}_{BM} = \sum_{i=1}^{B} \hat{\beta}_{iBM}$/B,

(ii) $SE(\hat{\beta}_{BM})$, the empirical Standard Error (SE) of β_{BM}over all simulations,

(iii) Design effect, defined as the ratio between the estimated variance of β_{BM} in each sampling design by the one in SRS (34),

(iv) Mean Square Error, MSE, given by $\left(\bar{\hat{\beta}}_{BM} - \beta_{BM}\right)^2 + \left(SE\left(\hat{\beta}_{BM}\right)\right)^2$,

(v) Coverage of the 95% confidence interval (CI) of β_{BM} and 95%CI length,

(vi) Power, number of times in which the null hypothesis ($\beta_{BM} = 0$) was rejected by the Wald test at 5% significance level in the two-phase Cox regression model.

All analyses were performed using `R software (version 3.5.2)` (35).

3. RESULTS

3.1 Design comparison

General results considering both common (~25% frequency) and rare biomarker (~5% frequency) are shown in **Table 1a-b**, respectively, under three censoring schemes (absent, low and high). Overall, the simulations showed that the β_{BM} was estimated without any noticeable bias for all designs. The standardized bias was always lower than 5% and the distributions of $\hat{\beta}_{BM}$ were symmetric for all sampling designs (Additional file Figure S1).

As shown in **Table 1a**, the PPS did not show much advantage respect to the SRS design. The empirical Standard Error of SRS and PPS were about the same, indicating no gain in efficiency. We found a small but not relevant increase of power in PPS stratified by the surrogate (2c in **Table 1a**) compared with traditional PPS.

On the other hand, the CC design improved power as compared with SRS reducing MSE and empirical Standard Error (for each scenario), with a further advantage when also the surrogate variable was used for stratification (3c in **Table 1a**). The stratification for risk factor and confounder (3a and 3b in **Table 1a**) showed a slight loss of efficiency with respect to the classical CC (3 in **Table 1a**).

When matching on time in the NCC and CM designs (4 and 5 in **Table 1**), a decrease of Mean Square Error and empirical Standard Error (increase in design effect) were obtained if compared with CC, PPS, and SRS in the presence of a censoring rate of ρ=0.4. CM presented higher design effect and shorter confidence intervals with respect to CC stratified by the surrogate for any censoring rate. Among all the considered designs, the CC stratified by the surrogate was showing the highest power. By note, in NCC and CM design, the final sample size of the second phase was sligthly lower than the fixed one (i.e. n=600, see **Table 1**) as resampling of controls is possible in these designs.

Similar performance results were obtained when a rare exposure (~5%) was considered (**Table 1b**). In general, with a rare exposure, performance of NCC and CM, in term of absolute bias, design effect and length of 95%CI, had an improvement with increasing censoring rates. The estimate of regression coefficient ($\hat{\beta}_{BM}$), length and coverage of its 95% confidence interval considering common and rare biomarker are given as Additional file Table S2.

Insert Table1

As sensitivity analysis, we evaluated the performance of the different designs modifying variables included in the weighted Cox model in (1). Results are presented in the Additional file (Table S3) and are consistent with previous results.

Interestingly, when the weighted Cox model was adjusted also for the risk factor variable (Table S3a), $X_{Risk\ Factor}$, there was an increase of power for all designs as compared with results of **Table 1a**. On the contrary, when the Cox model was adjusted for all variables available in our setting (i.e. confounder, X_{Conf} , risk factor $X_{Risk\ Factor}$ and surrogate, X_{Surr}, see Table S3b), power decreased.

3.2 Impact of surrogate

In this section, we explore the impact of the accuracy level of the surrogate variable in designs that use the surrogate as a stratification variable. As expected, higher sensitivity increased power and design effect in the CC design stratified by the surrogate and in the CM design (**Figure 3**). Interestingly, also results on the performance of post stratification by the surrogate variable (not used as strata in the design) showed an increase of power and design effect with an increase of the surrogate accuracy with a slightly lower performance than the CC design stratified by the surrogate.

Insert Figure 3

3.3 Power evaluation

In **Figure 4** we have explored the power by the size of the phase II sample in different designs. CM and NCC were more powerful up to a second phase sample size of nearly 500 individuals (1/4 of the entire cohort). For larger sample sizes, CC stratified for surrogate was the most powerful design. By note, both NCC and CM were sampled considering a 1:1 case-control ratio and controls could be resampled, thus the sample size of phase II was constrained to not exceed twice the number of events in the entire cohort.

Insert Figure 4

3.4 Application on the real data

The study that motivated our work was performed to evaluate the role of different genetic polymorphisms on treatment failure due to relapse (14, 15) and used data from a large Italian clinical trial (ClinicalTrials.gov identifier NCT00613457) (36). Clinical data and other information were available for the whole trial cohort (first phase) of 1999 consecutive patients newly diagnosed with childhood acute lymphoblastic leukemia between 2000 and 2006.

297 Biological samples were stored at diagnosis and were used to measure the genetic
298 polymorphism of interest (second phase). In the study of Franca et al. (15) the subsample on
299 which to measure the genetic polymorphism was chosen after classifying patients into six
300 strata according to the event of interest (relapse/no relapse) and a three-level risk group
301 stratification defined by prognostic features in the treatment protocol. Patients were sampled
302 at random without replacement from each stratum, according to an optimal sampling strategy
303 (37). In particular, the sampling fractions for each stratum were chosen proportionally to the
304 genetic variability reported within each of the strata to maximize the precision of the estimate
305 of the genotype effect on the outcome. Of note, this was possible only due to the availability
306 of pilot data on the genetic polymorphism of interest, that actually are not often available.
307 Overall, out of the 766 children for whom genotyping was required (approximately 1.5
308 controls for each case), the biomarker of interest (GST-θ) was obtained on 601 patients,
309 getting a hazard ratio (HR) of 1.34 (95%CI: 0.90-2.00). By breaking up the variance of the
310 coefficient of GST-θ into phase I and II contributions, we derived the efficiency of the design
311 with respect to the expected one in the full cohort (estimate of the minimum irreducible
312 uncertainty for the cohort) that resulted 54% by having genotyped 1/3 of the sample.
313 Interestingly, the efficiency we got was higher than the expected one in any of the CC designs
314 considered (see **Table 2**), as computed by simulations developed in this paper. Thus, the use
315 of pilot data for an optimal sampling strategy compensated the lack of a surrogate variable.

316 Power evaluation was not done in this study at the design stage, but according to our
317 simulations results, a sample size of n=601 subjects would have reached a power of 55% and
318 68% to detect an HR_{BM} of 1.3 and 1.5, respectively with a CC design. If CC stratified by
319 surrogate would have been considered, an increase of power would have been obtained (60%
320 and 71% respectively for an HR_{BM} of 1.3 and 1.5), but still not reaching a reasonable value
321 (i.e. 80%).

322 *Insert Table 2*

323 ## 4. DISCUSSION

324 This work underlined the importance of a careful study design in retrospective studies
325 evaluating a new research question using available cohort data on which to measure
326 additional characteristics, such as a new biomarker. It showed the advantages we can get in
327 terms of efficiency and power by using available data and the importance of power evaluation
328 in order to avoid useless studies. We also provide a tool to compute power by simulations.
329 From the simulation results, we found that the weighted Cox model provided valid estimates
330 of biomarker effect and good coverage probabilities in the considered designs. The
331 availability of auxiliary/surrogate variables of the biomarker of interest in phase I, the amount
332 of censoring and the prevalence of the biomarker, together with power considerations could
333 help researchers to identify the most efficient design. As expected, the CC provided better
334 efficiency with respect to the SRS design, while PPS did not show much advantage (5). If
335 some covariates are expected to be associated with the new variable of interest, it is
336 advantageous to use them to define strata in a two-phase design, especially if they have a
337 good accuracy in predicting it and in the presence of biomarkers with low frequency. Of note,
338 simulation results showed that using these surrogate variables of the biomarker just in the
339 analysis stage (and not in the stratified design) is also improving efficiency and power.
340 Interestingly, if a variable is associated both with the new variable and with the event of
341 interest, such as the “confounder”, or just with the event of interest, such as the “risk factor”,
342 using it to define strata did not show any advantage in power. Nevertheless, the inclusion of
343 the “risk factor” in the Cox model on the outcome is beneficial.
344 In the presence of censoring, sampling designs matching on time (NCC and CM) have shown
345 higher performance in term of design effect than CC and CC stratified by the surrogate
346 designs, respectively. Similar results were found by Borgan and Olsen, that also suggested to

combine the simple and counter-matching designs (sampling some controls by simple random sampling and others by stratified random sampling) (38). Higher design effect is not always followed by an improvement in power as the last one depends also on the direction of bias that actually is favouring the CC design (as shown in Additional file Table S2). Moreover, matched designs are constrained to have a fixed integer number for the case/control ratio and this could result as a disadvantage in some settings. In the absence of censoring, results showed that CC is more powerful compared to the all other designs. Stoer and colleagues have found similar results and have called this particular condition as "CC extreme" design (39), as in this setting controls have the longest possible follow-up (subjects event-free at the end of follow-up in the absence of censoring). We also found, similarly to (22, 26), that CM has a marked efficiency advantage especially when the biomarker is rare, as surrogate information help to sample subjects with the biomarker.

One limitation of our work is that we have considered only 1:1 matching ratio for NCC and CC designs, but we did a fair evaluation by comparing the perfomance of different designs at the same sample size. Moreover, to emulate the ALL data, we have considered only a moderate effect of the biomarker on the event and we have assumed time-constant coefficients. However the general indications coming from our work are consistent with previous results and among different setting, as well as for different spcification of the outcome model, as explored with sensitivity analysis. Moreover, the code developed, available at https://github.com/Fgraziano?tab=repositories, is helpful to investigate the power of different sampling designs in various setting.

CONCLUSIONS

Summarizing, for efficient selection of the subcohort, we recommend the use of the information available on the entire cohort as suggested in the flowchart of Additional file Figure S2 in suppoting information. If a surrogate variable of the biomarker of interest is

available, we suggest to use a case-control study stratified for the surrogate variable or a counter-matching design. The latter choice should be driven by the rate of censoring: if censoring is low we might opt for CC stratified, otherwise CM is more convenient. If the surrogate variable is not available, we should consider using a CC or a NCC design depending on the censoring rate. As NCC and CM designs are constrained by a fixed ratio between cases and controls (1:1, 1:2 …), the overall sampling fraction with respect to the number of cases should be also considered together with power evaluation. In fact, if sufficient power would be reached with a phase II sample size n of nearly all cases plus a number of controls equal to 1.5 the number of cases, a CC design would be more convenient with respect to the matched designs. Moreover, we found that optimal sampling strategies using potentially available pilot data are greatly efficient. Thus, optimal sampling strategies for survival data would be very useful together with an user-friendly instrument to compute power and efficiency for different sampling designs (also considering stratification variable) in the two-phase setting.

List of abbreviations: Acute Lymphoblastic Leukemia (ALL); Case-Control (CC), Counter-matching (CM), Hazard Ratio (HR), Nested Case-Control (NCC), Simple Random Sample (SRS), Probability Proportional to size (PPS)

DECLARATIONS

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable

CONSENT FOR PUBLICATION

Not applicable

AVAILABILITY OF DATA AND MATERIALS

The simulation codes that support the findings of this study are openly available at https://github.com/Fgraziano?tab=repositories.

COMPETING INTERESTS

The authors declare that they have no competing interests

FUNDING

This work was supported by the Italian Ministry of Education, University and Research [grant SIR-MIUR RBSI14LOVD, 2017-NAZ-0446 (PRIN 2017) and Department of Excellence project PREMIA (PREcision MedIcine Approach: bringing biomarker research to clinic)]. The sponsor has no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

AUTHORS' CONTRIBUTIONS

All authors commented on the draft and the interpretation of the findings, read and approved the final manuscript.

FG: drafting the article, data analysis and interpretation, critical revision of the article, final approval of the version to be published.

PR: conception the work (PI), supervision of the data collection, data analysis and interpretation, drafting the article, critical revision of the article, final approval of the version to be published.

MGV: critical revision of the article, final approval of the version to be published.

ACKNOWLEDGMENT

We thank the Italian Ministry of Education, University and Research

REFERENCES

1. Kannel WB. Bishop lecture. Contribution of the Framingham Study to preventive cardiology. J Am Coll Cardiol. 1990;15(1):206-11.

2. Buist AS. The US Lung Health Study. Respirology. 1997;2(4):303-7.

3. Evans A, Salomaa V, Kulathinal S, Asplund K, Cambien F, Ferrario M, et al. MORGAM (an international pooling of cardiovascular cohorts). Int J Epidemiol. 2005;34(1):21-7.

4. Lee J-E. How Should Biobanks Collect Biosamples for Clinical Application? A 20-year Biomarker-related Publication and Patent Trend Analysis. Osong public health and research perspectives. 2018;9(3):105-11.

5. Borgan Ø. Cohort sampling in epidemiological studies. Preprint series Statistical Research Report http://urn nb no/URN: NBN: no-23420. 2001.

6. Borgan Ø, Samuelsen SO. A review of cohort sampling designs for Cox's regression model: Potentials in epidemiology. Norsk Epidemiologi. 2003;13(2).

7. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. Am J Epidemiol. 2009;169(11):1398-405.

8. Lumley T. Complex surveys: a guide to analysis using R: John Wiley & Sons; 2011.

9. Langholz B. Use of cohort information in the design and analysis of case-control studies. Scandinavian Journal of Statistics. 2007;34(1):120-36.

10. Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. American journal of epidemiology. 1990;131(1):169-76.

11. Cai J, Zeng D. Sample size/power calculation for case–cohort studies. Biometrics. 2004;60(4):1015-24.

12. Haneuse S, Saegusa T, Lumley T. osDesign: An R Package for the Analysis, Evaluation, and Design of Two-Phase and Case-Control Studies. J Stat Softw. 2011;43(11).

13. Rivera-Rodriguez C, Spiegelman D, Haneuse S. On the analysis of two-phase designs in cluster-correlated data settings. Stat Med. 2019;38(23):4611-24.

14. Rebora P, Valsecchi MG. Survival estimation in two-phase cohort studies with application to biomarkers evaluation. Stat Methods Med Res. 2016;25(6):2895-908.

15. 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.

16. Marubini E, Valsecchi MG. Analysing survival data from clinical trials and observational studies: John Wiley & Sons; 2004.

17. Rosén B. On sampling with probability proportional to size. Journal of Statistical Planning and Inference. 1997;62(2):159-91.

18. Vandenbroucke JP, Pearce N. Case-control studies: basic concepts. Int J Epidemiol. 2012;41(5):1480-9.

19. Haneuse S, Rivera-Rodriguez C. On the analysis of case–control studies in cluster-correlated data settings. Epidemiology. 2018;29(1):50-7.

20. Delcoigne B, Stoer NC, Reilly M. Valid and efficient subgroup analyses using nested case-control data. Int J Epidemiol. 2018.

21. Ernster VL. Nested case-control studies. Prev Med. 1994;23(5):587-90.

22. Langholz B, Clayton D. Sampling strategies in nested case-control studies. Environ Health Perspect. 1994;102 Suppl 8:47-51.

23. Samuelsen SO. A psudolikelihood approach to analysis of nested case-control studies. Biometrika. 1997;84(2):379-94.

24. Samuelsen SO, Ånestad H, Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. Scandinavian journal of statistics. 2007;34(1):103-19.

25. Rivera C, Lumley T. Using the whole cohort in the analysis of countermatched samples. Biometrics. 2016;72(2):382-91.

26. Cologne JB, Sharp GB, Neriishi K, Verkasalo PK, Land CE, Nakachi K. Improving the efficiency of nested case-control studies of interaction by selecting controls using counter matching on exposure. Int J Epidemiol. 2004;33(3):485-92.

27. Steenland K, Deddens JA. Increased precision using countermatching in nested case-control studies. Epidemiology. 1997;8(3):238-42.

28. Lin DY. On fitting Cox's proportional hazards models to survey data. Biometrika. 2000;87(1):37-47.

29. Binder DA. Fitting Cox's proportional hazards models from survey data. Biometrika. 1992;79(1):139-47.

30. Laitila T, Olofsson J. A two-phase sampling scheme and πps designs. Journal of Statistical Planning and Inference. 2011;141(5):1646-54.

31. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006;25(24):4279-92.

32. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Stat Med. 2005;24(11):1713-23.

33. Rebora P, Antolini L, Glidden DV, Valsecchi MG. Crude incidence in two-phase designs in the presence of competing risks. BMC Med Res Methodol. 2016;16:5.

34. Kish L. Survey Sampling Wiley. New York. 1965.

35. TeamR RC. A Language and Environment for Statistical Computing. 2015. R Foundation for Statistical Computing, Vienna, Austria:[URL https://www.R-project.org/] Accessed November. 2018;26.

36. Moricke A, Zimmermann M, Valsecchi MG, Stanulla M, Biondi A, Mann G, et al. Dexamethasone vs prednisone in induction treatment of pediatric ALL: results of the randomized trial AIEOP-BFM ALL 2000. Blood. 2016;127(17):2101-12.

494 37. Reilly M. Optimal sampling strategies for two-stage studies. American journal of
495 epidemiology. 1996;143(1):92-100.
496 38. Borgan O, Olsen EF. The Efficiency of Simple and Counter-matched Nested Case-
497 control Sampling. Scandinavian journal of statistics. 1999;26(4):493-509.
498 39. Støer N, Salim A, Bokenberger K, Karlsson I, Reilly M. Is the matched extreme case–
499 control design more powerful than the nested case–control design? Statistical Methods in
500 Medical Research.0(0):0962280218778624.

501

Table 1. Bias, empirical standard error, mean square error, power and design effect of the biomarker regression coefficient estimate ($\hat{\beta}_{BM}$) for the full cohort and different sampling designs. Accuracy of surrogate: sensitivity (i.e. probability of having a positive surrogate if the biomarker is positive)=0.7 and specificity (i.e. probability of having a negative surrogate if the biomarker is negative)=0.7, biomarker common (a) and rare (b)

a) SAMPLING DESIGN	Stratification Variable	n	BIAS			SE EMPIRICAL			MSE			POWER (%)			DESIGN EFFECT		
			Censoring Rate			Censoring Rate			Censoring Rate			Censoring Rate			Censoring Rate		
			0	0.1	0.4	0	0.1	0.4	0	0.1	0.4	0	0.1	0.4	0	0.1	0.4
Full cohort	-	**2000**	0.008	-0.015	0.009	0.093	0.095	0.112	0.009	0.009	0.013	99	97	95	-	-	-
1. SRS	-	**600**	0.004	-0.013	0.006	0.182	0.187	0.206	0.033	0.035	0.042	64	58	53	-	-	-
2. PPS	Event	**599**	0.007	-0.015	0.007	0.173	0.180	0.199	0.029	0.033	0.039	65	58	54	1.003	1.003	1.005
2a. PPS	Event; Risk factor	**598**	0.008	-0.016	0.004	0.172	0.175	0.205	0.029	0.031	0.042	65	58	52	1.002	1.003	1.002
2b. PPS	Event; Confounder	**598**	0.003	-0.015	0.002	0.174	0.179	0.203	0.030	0.032	0.041	65	57	51	0.999	1.002	1.000
2c. PPS	Event; Surrogate	**598**	0.007	-0.013	0.013	0.161	0.171	0.190	0.026	0.029	0.036	69	64	57	1.106	1.129	1.104
3. CC	Event	**600**	0.011	-0.008	0.019	0.159	0.158	0.179	0.025	0.025	0.032	74	68	67	1.179	1.219	1.352
3a. CC	Event; Risk factor	**600**	0.010	-0.009	0.008	0.162	0.166	0.182	0.026	0.028	0.033	72	65	62	1.139	1.176	1.307
3b. CC	Event; Confounder	**600**	0.012	-0.015	0.010	0.162	0.161	0.175	0.026	0.026	0.031	73	65	66	1.182	1.187	1.354
3c. CC	Event; Surrogate	**600**	0.008	-0.016	0.012	0.148	0.153	0.170	0.022	0.024	0.029	76	71	69	1.334	1.363	1.495
4. NCC	Event	**550**	0.008	-0.018	0.014	0.169	0.165	0.175	0.029	0.028	0.031	68	63	67	1.066	1.144	1.378
5. CM	Event; Surrogate	**546**	-0.044	-0.058	-0.009	0.151	0.153	0.165	0.025	0.027	0.027	67	61	67	1.379	1.395	1.536

b) SAMPLING DESIGN	Stratification Variable	n	BIAS			SE EMPIRICAL			MSE			POWER (%)			DESIGN EFFECT		
			Censoring Rate			Censoring Rate			Censoring Rate			Censoring Rate			Censoring Rate		
			0	0.1	0.4	0	0.1	0.4	0	0.1	0.4	0	0.1	0.4	0	0.1	0.4
Full cohort	-	**2000**	-0.028	-0.004	-0.023	0.193	0.199	0.225	0.038	0.039	0.052	52	53	41	-	-	-
1. SRS	-	**600**	-0.045	-0.041	-0.093	0.376	0.392	0.802	0.144	0.156	0.652	23	23	19	-	-	-
2 PPS	Event	**599**	-0.051	-0.023	-0.060	0.380	0.394	0.453	0.147	0.155	0.209	25	24	19	1.079	1.131	1.139
2a. PPS	Event; Risk factor	**598**	-0.048	-0.039	-0.081	0.372	0.393	0.721	0.141	0.156	0.526	22	24	21	1.085	1.109	1.145
2b. PPS	Event; Confounder	**598**	-0.057	-0.033	-0.059	0.381	0.393	0.453	0.148	0.156	0.209	23	23	20	1.077	1.104	1.140
2c. PPS	Event; Surrogate	**598**	-0.055	-0.022	-0.073	0.388	0.391	0.597	0.153	0.153	0.362	24	25	19	1.090	1.147	1.158
3. CC	Event	**600**	-0.003	0.021	0.003	0.332	0.339	0.368	0.110	0.116	0.135	24	26	22	1.257	1.317	1.509
3a. CC	Event; Risk factor	**600**	-0.011	0.015	0.006	0.345	0.357	0.384	0.119	0.127	0.148	23	24	21	1.193	1.283	1.418
3b. CC	Event; Confounder	**600**	-0.023	0.024	0.011	0.328	0.344	0.363	0.108	0.119	0.132	23	26	22	1.267	1.329	1.513
3c. CC	Event; Surrogate	**565**	-0.018	0.011	-0.010	0.311	0.313	0.345	0.097	0.098	0.119	26	29	24	1.419	1.607	1.757
4. NCC	Event	**545**	-0.020	0.024	-0.000	0.354	0.355	0.363	0.126	0.127	0.132	21	24	21	1.126	1.222	1.513
5. CM	Event; Surrogate	**529**	-0.057	-0.041	-0.028	0.311	0.315	0.341	0.099	0.101	0.117	26	27	25	1.541	1.679	1.823

Legend: SRS–Simple Random Sample; PPS- Probability Proportional to size; CC- Case-Control; NCC- Nested Case-Control; CM- Counter-matching

Table 2. Efficiency (refers to the full cohort), design effect (refers to Simple Random Sampling) and power for Case-Control (CC) designs with hypothetical hazard ratio of the biomarker of interest (HR_{BM}) of 1.3 and 1.5, biomarker common (25%), censoring rate ρ=0.1, type I error 0.05

	EFFICIENCY		
	Case-control	**CC stratified by surrogate**	**CC stratified by risk factor**
$\boldsymbol{HR_{BM}}$ **= 1.3**	38.91%	43.06%	34.47%
$\boldsymbol{HR_{BM}}$ **= 1.5**	36.26%	38.51%	32.73%
	DESIGN EFFECT		
	Case-control	**CC stratified by surrogate**	**CC stratified by risk factor**
$\boldsymbol{HR_{BM}}$ **= 1.3**	1.23	1.37	1.20
$\boldsymbol{HR_{BM}}$ **= 1.5**	1.22	1.36	1.18
	POWER		
	Case-control	**CC stratified by surrogate**	**CC stratified by risk factor**
$\boldsymbol{HR_{BM}}$**= 1.3**	54.8%	60.1%	54.3%
$\boldsymbol{HR_{BM}}$**= 1.5**	68.1%	70.65%	65.4%

LIST OF FIGURES

Figures

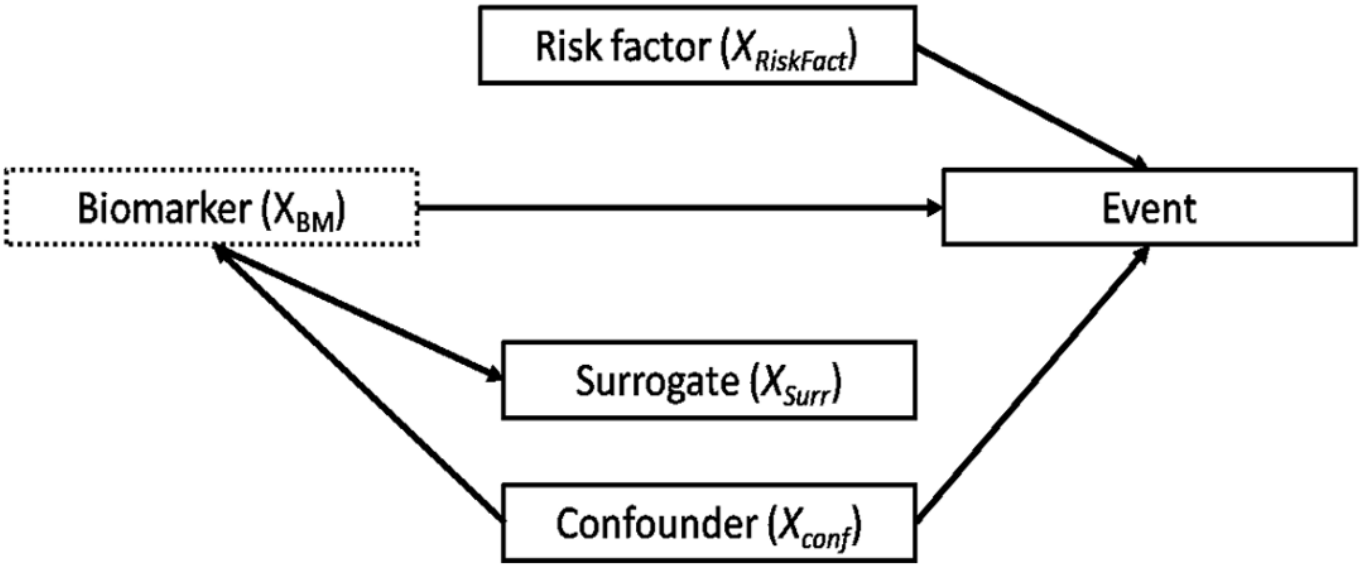

Figure 1

Causal diagram where the variables in the boxes are connected each other through the black arrows, denoting association. The dashed line box indicates a variable measured only in the sub-cohort

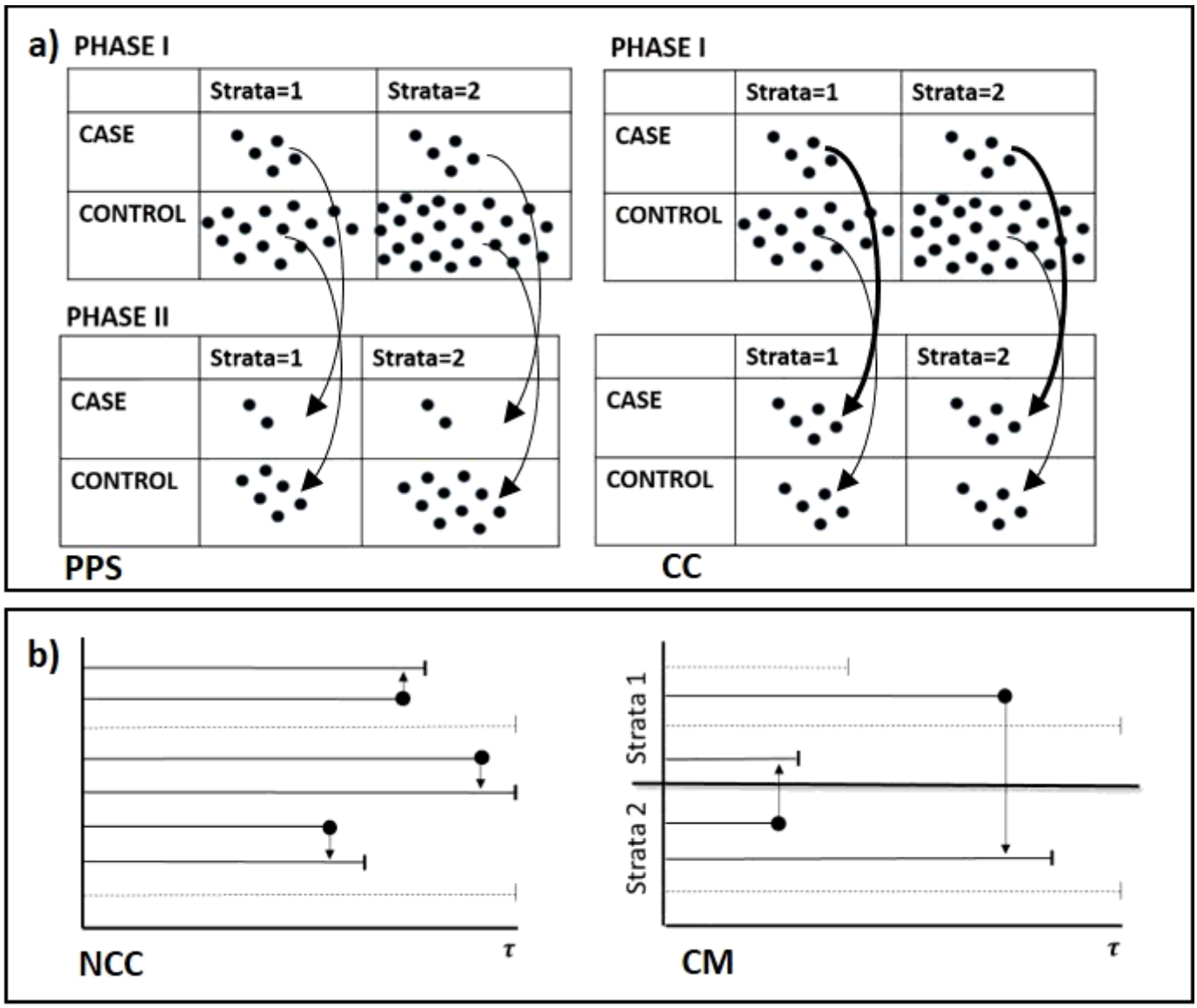

Figure 2

Probability Proportional to size (PPS) and Case-Control (CC) sampling from phase I cohort are shown in the upper part (panel a), left and right, respectively. Dots represents individuals in the strata (case or control and strata=1 or strata=2). Arrows correspond to the sampling from phase I to phase II. The number of sampled individuals in each stratum (phase II) depends on the sampling design. Nested Case-Control (NCC) and Counter-matching (CM) sampling are shown below in the figure (panel b). The lines represent the follow-up over which individuals are observed and the solid lines represented the sampled subjects. Black dot symbol represents the occurrence of an event and the arrow indicated the corresponding sampled control. For NCC, sampling is conducted in the same stratum and for CM, cases are matched with controls from the opposite stratum

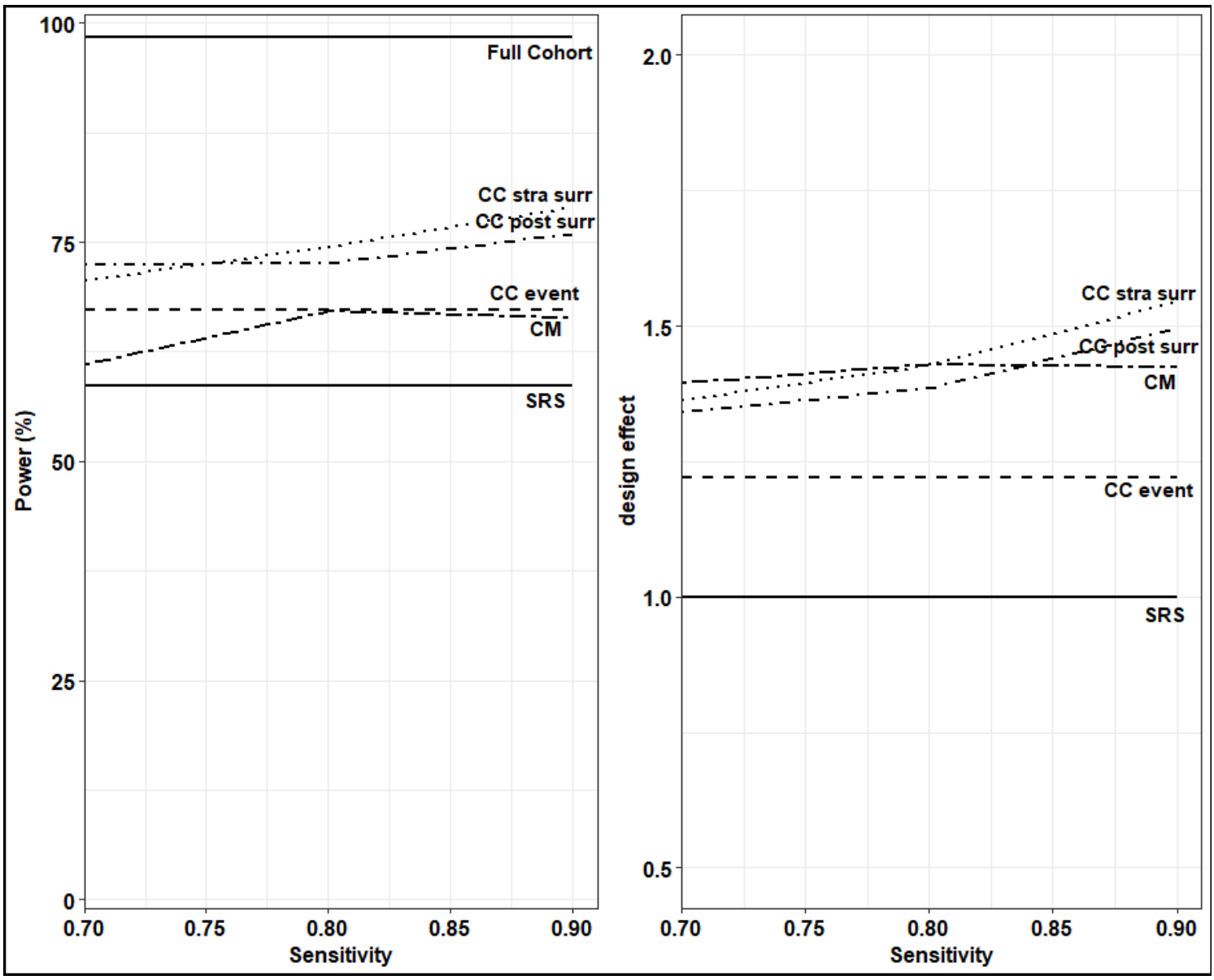

Figure 3

Power and design effect for different sensitivity levels (i.e. probability of having a positive surrogate if the biomarker is positive) of the surrogate variable. Scenario: specificity (i.e. probability of having a negative surrogate if the biomarker is negative) =0.7, censoring rate ρ=0.1, hazard ratio of biomarker =1.5 and sample size of second phase (n) =600. Legend: CC stra surr (Case-Control stratified by surrogate), CC post surr (Case-Control post stratified by surrogate), CC event (Case-Control), CM (Counter-Matching) and SRS (simple random sampling)

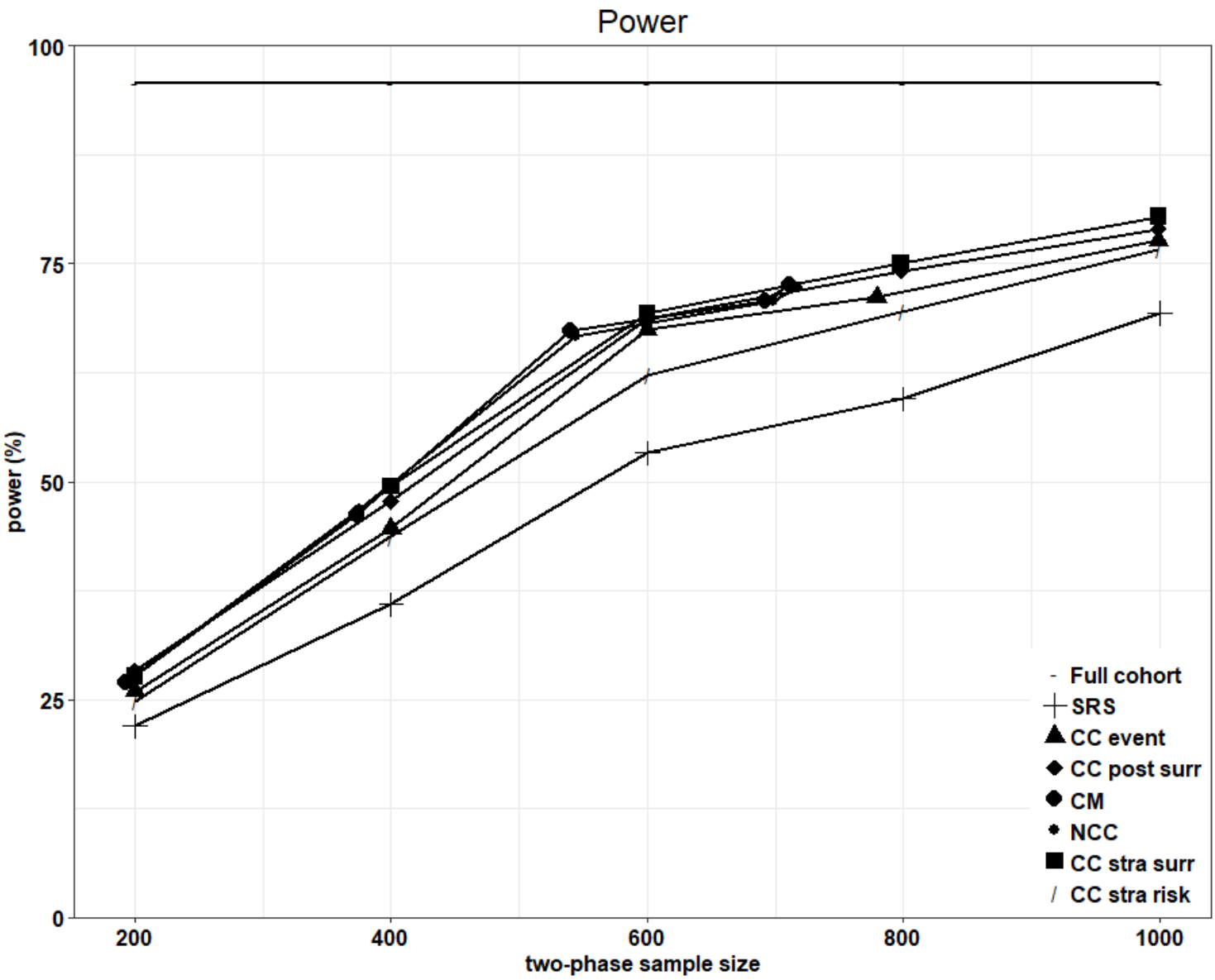

Figure 4

Power for different sample sizes of second phase (n). Scenario: censoring rate ρ=0.4, common biomarker (25%), hazard ratio of biomarker=1.5, sensitivity (i.e. probability of having a positive surrogate if the biomarker is positive) =0.7 and specificity (i.e. probability of having a negative surrogate if the biomarker is negative)=0.7. Legend: CC stra surr (Case-Control stratified by surrogate), CC post surr (Case-Control post stratified by surrogate), CC event (Case-Control), CM (Counter-Matching) and SRS (simple random sampling)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile.pdf