Analysis of Clinical Cohort Data Using Nested Case-control and Case-cohort Sampling Designs

A Powerful and Economical Tool

K. Ohneberg1,2; M. Wolkewitz1,3; J. Beyersmann3; M. Palomar-Martinez4,5; P. Olaechea-Astigarraga6; F. Alvarez-Lerma7; M. Schumacher1

1Institute for Medical Biometry and Statistics, Medical Center University of Freiburg, Freiburg, Germany;
2Freiburg Center for Data Analysis and Modelling, University of Freiburg, Freiburg, Germany;
3Institute of Statistics, Ulm University, Ulm, Germany;
4Hospital Universitari Arnau de Vilanova, Lleida, Spain;
5Universitat Autònoma de Barcelona, Barcelona, Spain;
6Service of Intensive Care Medicine, Parc de Salut Mar, Barcelona, Spain;
7Service of Intensive Care Medicine, Hospital de Galdakao-Usansolo, Bizkaia, Spain;

E-mail: ohneberg@imbi.uni-freiburg.de

Methods Inf Med 2015; 54: 505–514
http://dx.doi.org/10.3414/ME14-01-0113
received: November 6, 2014
accepted: April 13, 2015
epub ahead of print: June 25, 2015

1. Introduction

Cox’ proportional hazards (PH) model is widely used in order to analyze data of cohort studies. For that purpose information on exposure to a potential risk factor, as well as type and time of an event is required for every patient within the cohort. The observations could be subject to independent left-truncation and right-censoring. In the situation of a large cohort in which the event of interest is rare and interest is in estimating the effect of a certain risk factor, that is rather elaborate or expensive to obtain, an attractive alternative is given by some cohort sampling designs. One of them is the well-known nested case-control (NCC) design, a case-control study embedded into an existing cohort. Instead of taking all individuals who are still at risk of experiencing an event, one samples a generally fixed number of controls from those at risk at each event time. We consider the simple nested case-control design as proposed by Liddell et al. [1]. If some additional information is available for the full cohort it is possible to apply a stratified NCC design incorporating such information in the sampling design, called counter-matching [2]. Another sampling approach is a case-cohort (CC) design, where the sample consists of a random subcohort taken at the outset of the study from the entire cohort plus all the cases that occur outside of the subcohort. Sampling designs like the NCC or CC design
allow to strongly reduce the number of patients for whom covariate information is required, a major benefit if the potential risk factor is costly or complex to obtain. Schoenfeld [3] proposed a sample size formula for the log rank test and later on showed that this sample size formula is valid also for Cox’ PH model [4]. In the meantime this formula is widespread and fundamental for sample size calculation for survival data and several extensions are developed. For an overview see for example the chapter on Sample Size Calculations for Clinical Trials [5] in The Handbook of Survival Analysis [6]. We find that the power formulas for the NCC design as well as for the CC design are also closely linked to Schoenfeld’s formula.

2. Objectives

Our aim is to close the gap between study designs of epidemiological studies where sampling designs are widely used and clinical studies where sampling strategies are less frequently applied. We want to compare the results when analyzing a large clinical cohort using sampling designs instead of a random subsample or the full cohort with respect to savings in the number of patients for whom covariate information is required and loss in precision coming along with the reduced data set. In choosing between the nested case-control design and the case-cohort design we want to depict some assets and drawbacks coming along with each sampling design and go into some practical considerations.

3. Methods

3.1 Nested Case-control Design

The nested case-control (NCC) design was introduced by Thomas in the addendum to Liddell et al. [1], furthermore see Prentice and Breslow [7]. The NCC design actually yields a case-control study embedded in an existing cohort. Incidence density sampling [8, 9] should be applied, where controls for each case are randomly selected from the risk set corresponding to the case. In a NCC design with a \(1:m\) sampling at each event time \(m\) controls are randomly selected from those individuals still at risk. In order to derive unbiased estimators it is important that the same subject may serve as a control for more than one case. Hence if an individual is at risk at multiple event times, at every observed failure time this individual could be sampled as control.

Thus sampling of controls is done with replacement with respect to different event times. In contrast, the sampling of \(m\) controls per case is carried out without replacement, so that effectively \(m\) different controls are available per case. Another essential issue is that an individual can be sampled as control before itself becoming a case at a later time. Being aware that the incidence density sampling is conducted in order to generate a control group that would be comparable to the entire risk set used for analysis in the partial likelihood of the full cohort, the above mentioned requirements seem reasonable.

Exemplarily we plotted the observation time of 10 patients admitted to an intensive care unit (ICU), who either experienced the event of interest or are censored (Figure 1). In the upper part of the figure we display the NCC sampling. The vertical lines represent the risk set corresponding to each case. Assuming a \(1:1\) sampling the individuals marked with \(x\) could randomly be sampled as controls. Note that patient 10 is sampled twice and patient 2 who experienced the event at day 8 is sampled as a control for another patient with an event at day 4.

The partial likelihood for a NCC design resembles the full cohort partial likelihood except that for every failure time the sum in the denominator includes only the case and the sampled controls instead of all individuals who were still at risk just before the event. Breslow [10] points out that the NCC partial likelihood corresponds to a conditional likelihood for matched case-control data under a logistic regression model what “further strengthens the notion that one is estimating the same parameters in cohort studies and case-control studies”. For asymptotic consistency and normality of relative risks estimated by partial likelihood under NCC sampling see Goldstein and Langholz [11].

3.2 Case-cohort Design

The case-cohort (CC) design for failure time data was introduced by Prentice [12]. Within this design a random sample of the entire cohort, the so called subcohort, serves as source of controls. For analysis then all the individuals with the event of interest are included, additionally censored observations are included but only if they occur in the subcohort. That means covariate information is required for the subcohort plus all the cases that occur outside of the subcohort.

In Figure 1 (lower part) we randomly sampled five patients for the subcohort, marked by the horizontal line. Within the subcohort one patient experienced the event of interest (patient 2). For analysis the patients sampled to the subcohort plus three cases that occur outside of the subcohort are considered.

Prentice [12] proposed a pseudo-likelihood approach with the risk set at each event time consisting of the patients in the subcohort who are still at risk for an event and the individual with the event at that time point. An individual not sampled for the subcohort with an event does not enter the risk sets earlier in time but is considered to be at risk just before its own failure time. Otherwise one would use information about the future what might lead to biased estimators [13].

In order to estimate the hazard ratio \(\exp(b)\) (with \(b\) the unknown regression coefficient) using a Cox regression model in the full cohort, Cox proposed to maximize a partial likelihood [14, 15]. For a detailed insight and some extensions see for example the chapter Cox Regression Model in The Handbook of Survival Analysis [6]. Using a CC design instead of the full cohort results in the score contributions (the first derivative of the likelihood with respect to the unknown regression coefficient) being no longer independent, thus a partial likelihood can not be applied. Instead, a pseudo-likelihood corresponding to a weighted Cox regression model is used with weights proposed for example by Prentice [12], Self and Prentice [16] and Barlow [17]. The weights are required in order to adjust for the reduced risk set within the likelihood. An overview on dif-
Different weighting methods is given by Barlow et al. [13] and more recently by Petersen et al. [18] and Onland-Moret et al. [19]. As an extension Breslow et al. [20] proposed to use information available for the full cohort to improve the efficiency of the case-cohort design.

### 3.3 Simple Random Sampling

Simple random sampling provides a rather pragmatic approach in the situation of a large cohort when it is too costly to obtain covariate information for all the patients. Thereby a random subset of the entire cohort is collected without considering if a patient experienced the event of interest or not (i.e. without consideration of case or control status). Analysis then is performed within the random subset, that means covariate information is required only for the patients sampled. Broadly speaking, the simple random sampling complies with the CC design if analysis is performed only in the subcohort, i.e. without adding the patients with the event of interest outside of the subcohort.

Indeed we found that simple random sampling is applied in practice, see for example the meta-analysis of Schöttker et al. [21]. In addition to studies using the full cohort or a nested case-control design also studies based on a random subsample, equivalent to a simple random sampling approach, are included.

In our data example we will compare the simple random sampling to the results obtained by a NCC or CC design or using the full cohort.

### 3.4 Study Design and Power

Schoenfeld [3] proposed a formula for the power of a log rank test that turned out to be valid also for Cox’ PH model [4]. For a given number of observed events $N$, the proportion of patients exposed to the risk factor, an assumed hazard ratio $\theta$ and a level of significance $\alpha$ the power of Cox’ PH model for analyzing the full cohort is given by \[\text{Equation 1.}\]

Schoenfeld’s formula can also be used for calculating the power of a random sampling approach by adjusting for the reduced number of observed events. Assume that $\pi$ denotes the fraction of patients sampled to the random subset, then the expected number of observed events is given by $N\pi$. The power can then be calculated by \[\text{Equation 2.}\]

Lachin [22] derived a sample size formula for a conditional logistic regression model in a multiply matched case-control study design. It turns out that the power of a NCC design is given by Schoenfeld’s formula by adding a multiplication factor that only depends on the number of controls matched to each case (\[Equation 3.\]). This factor $m/(m + 1)$ was first published by Ury [23] for the asymptotic relative efficiency of a NCC sampling relative to the full cohort under $H_0: \beta = 0$, that is a hazard ratio $\exp(\beta)$ equal one meaning no effect of the covariate. Pang [24] stated that there seems to be a “traditional dictum” that four controls per case are necessary to achieve results that are similar to those given by the full cohort. This statement seems reasonable taking Lachin’s formula (\[Equation 3.\]) into account, as the choice of four controls matched to each case leads to a multiplication factor of about 0.9 and thus power results are comparable to the full cohort.

For the case-cohort design Cai and Zeng [25] derived a formula for the power assuming a single binary exposure variable, equal censoring distributions in both groups, a low event rate in the full cohort and no tied event times. Their power formula (\[Equation 4.\]) complies with Schoenfeld’s formula for the full cohort with a factor that only depends on the number of observed events $N$, the subcohort size $\hat{n}$ and the full cohort size $n$. If we compare the power of a simple random
sampling approach (\ref{eq:power_ncc}) to the power of a CC design (\ref{eq:power_cc}) and assume that the sampling fraction is sampled it becomes obvious that the power is decreased considerably choosing a simple random sampling instead of a CC design as

$$\pi \sim \frac{\hat{n}}{\bar{n} - \hat{n}}.$$ 

3.5 ICU Data

Our data example \cite{26} constitutes of 6567 patients admitted to a Spanish intensive care unit (ICU). Interest is in risk factors for nosocomial infections, which are infections acquired during a hospital-stay. A total of 432 patients (6.58\%) acquired an infection during ICU stay. For illustration we choose the APACHE (acute physiology and chronic health evaluation) score as risk factor of interest, as this score is available for all patients and therefore allows us to compare the results of the sampling designs to the results obtained by analyzing the full cohort. The APACHE score is a severity-of-disease classification system and assumed to have an influence on the incidence of infection. A higher APACHE score corresponds to a more severe disease and therefore to a higher risk of death. For illustration we consider a dichotomized APACHE-Score, where patients with an APACHE-Score greater than 15 (2104 patients) are considered as being exposed. We are interested in the cause-specific hazard for infection. End of the ICU stay and death constitute competing events for the occurrence of infection and are hence technically treated as censored.

3.6 Sampling Methods

We analyzed the ICU data set using a NCC, CC or simple random sampling approach and compared the results to those obtained by analyzing the full cohort. For each sampling we calculated the estimated log hazard ratio $\hat{\beta}$, the corresponding standard error $SE$ and the 95\% confidence interval. In order to quantify the savings in resources we also report the mean number of distinct patients required for the respective design. A total of 432 patients acquired an
infection, therefore we have 432 cases for the NCC and CC design respectively. For the NCC design we considered a varying number of \( m \in \{1, 2, 3, 4\} \) controls matched to each case. For the CC design we used subcohort sizes equal to the number of controls within the NCC design times the number of the cases. Our intention of this choice of subcohort sizes was to be able to compare the results of the NCC to the CC design with respect to the mean number of distinct persons included. With respect to the choice of the subcohort size, see also Kulathinal et al. [31], who report their experiences with the CC design for analyzing the multinational collaborative MORGAM study. They especially focus on the design of a CC study and propose a subcohort size of twice the number of observed events. Note that this approach is possible for retrospectively conducted CC studies, as the number of the cases has to be known at the time of sampling. For CC analysis we chose weights as initially proposed by Prentice [12], as in comparisons of different weighting methods Prentice’s weights were found to perform very well [19; 18]. Yet note that we aim at estimating relative risks in terms of hazard ratios, if one is interested in estimating absolute risks the weights proposed by Barlow might be the first choice. In order to keep our overview simple (Table 1), we decided to only show results using Prentice’s weights for the CC design.

We also used the simple random sampling approach with random subset sizes equal to the mean number of distinct persons of the NCC design. Whereas the NCC and CC approach include all the 432 cases for analysis, the random sampling approach only includes patients with an event that are randomly selected to the subset. Therefore we also report the mean number of events included for the RS approach.

### 3.7 Comparing Sampling Methods

In order to compare the results of various sampling methods with the results by analyzing the full cohort we choose the bootstrap method. We generated 100 independent bootstrap samples with replacement from the full cohort with equal total number of patients in each bootstrap sample (6567 patients). For each bootstrap sample we then analyze the full cohort and estimate the log hazard ratio (\( \hat{\beta} \)), the standard error and the 95% confidence interval. Note that the number of events (in the original data set 432 infections) varies among the bootstrap samples. For every bootstrap sample we then choose a NCC, CC and RS approach and again calculated the log hazard ratio, standard error and 95% confidence interval. In a final step we calculated the mean values out of 100 bootstrap samples for the various designs.

### 4. Results

As information within our ICU data example is available for the full cohort we can compare the results of Cox’ PH model in the full cohort with those obtained by using random sampling, NCC or CC approach. The results of our data analysis is given in Table 1. We generated 100 bootstrap samples of the full cohort and for each design present the respective mean values.

Analysis of the full cohort reveals an estimated regression coefficient \( \hat{\beta} = 0.462 \) with standard error \( \hat{SE} = 0.103 \) and 95% confidence interval \([0.260, 0.665]\). For this result covariate information is required for all 6567 patients within the cohort.

Using a NCC design with a varying number of \( m \in \{1, 2, 3, 4\} \) controls matched to each case, the standard error decreases from 0.143 to 0.115 and the estimated regression coefficient approximates the parameter obtained by analyzing the full cohort \( \hat{\beta} = 0.462 \). While the standard errors are only slightly greater, we observe a considerable reduction in the number of distinct patients required. For example for the 1:3 sampling covariate information of 1239 patients is required instead of 6567 patients for the full cohort.

For the CC design with varying subcohort sizes we get slightly greater standard errors compared to the NCC design. For example using a subcohort of 864 patients yields a mean standard error of 0.162 compared to a NCC design with a 1:3 sampling leading to a mean standard error of 0.119 for an almost equal number of distinct persons (about 1240 patients) required. Nevertheless the mean parameter estimates are all in all fitting well to the one obtained in the full cohort.

### Table 1

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\beta} )</th>
<th>( \hat{SE} )</th>
<th>95% CI</th>
<th>Mean(dist.p)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full cohort</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 : 1 sampling</td>
<td>0.444</td>
<td>0.143</td>
<td>[0.163, 0.724]</td>
<td>761</td>
</tr>
<tr>
<td>1 : 2 sampling</td>
<td>0.454</td>
<td>0.124</td>
<td>[0.211, 0.698]</td>
<td>1023</td>
</tr>
<tr>
<td>1 : 3 sampling</td>
<td>0.456</td>
<td>0.119</td>
<td>[0.224, 0.688]</td>
<td>1239</td>
</tr>
<tr>
<td>1 : 4 sampling</td>
<td>0.457</td>
<td>0.115</td>
<td>[0.231, 0.683]</td>
<td>1426</td>
</tr>
<tr>
<td><strong>CC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subcohort size</td>
<td>432 patients in subcohort</td>
<td>0.473</td>
<td>0.209</td>
<td>[0.064, 0.881]</td>
</tr>
<tr>
<td></td>
<td>864 patients in subcohort</td>
<td>0.447</td>
<td>0.162</td>
<td>[0.130, 0.763]</td>
</tr>
<tr>
<td></td>
<td>1296 patients in subcohort</td>
<td>0.461</td>
<td>0.143</td>
<td>[0.181, 0.740]</td>
</tr>
<tr>
<td></td>
<td>1728 patients in subcohort</td>
<td>0.451</td>
<td>0.132</td>
<td>[0.192, 0.709]</td>
</tr>
<tr>
<td><strong>RS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random subset size = Mean(dist.p) for NCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>761 patients (mean 50 events)</td>
<td>0.458</td>
<td>0.310</td>
<td>[–0.150, 1.066]</td>
<td>761</td>
</tr>
<tr>
<td>1023 patients (mean 68 events)</td>
<td>0.453</td>
<td>0.266</td>
<td>[–0.069, 0.975]</td>
<td>1023</td>
</tr>
<tr>
<td>1239 patients (mean 82 events)</td>
<td>0.488</td>
<td>0.242</td>
<td>[0.013, 0.964]</td>
<td>1239</td>
</tr>
<tr>
<td>1426 patients (mean 94 events)</td>
<td>0.459</td>
<td>0.224</td>
<td>[0.020, 0.898]</td>
<td>1426</td>
</tr>
</tbody>
</table>
A considerable loss in precision is observed by using the random sampling approach: While incorporating equal numbers of distinct persons, the standard errors are notably greater and accompanying confidence intervals wider.

The histograms of the estimated regression coefficient in 100 bootstrap samples using the simple random sampling, CC and NCC design are given in Figure 4 and clearly emphasize the high variability within the random sampling design compared to a CC or NCC design. The Random Sampling (RS) with 1239 patients, the CC design with 864 patients sampled to the subcohort and the NCC design with 1:3 sampling yield almost the same number of distinct persons (about 1240 patients). When looking at the respective histograms in Figure 4 the RS clearly shows the greatest variability, followed by the CC and the NCC, the latter with the narrowest histogram, yet being closest to the estimated regression coefficient in the full cohort.

We also calculated the power for analyzing our ICU data set according to the power formulas (Equations 1–4) for the different study designs (Figure 2). Using the full cohort with 6567 patients yields a power of approximately 97%. We plotted the power for the different sampling design with respect to the mean number of distinct persons involved (first x-axis). The CC and NCC design yield about the same power with respect to equal numbers of distinct persons required. For example the NCC design with a 1:3 sampling corresponding to 1239 patients included yields a power of approximately 90% just like a CC design with a sampling fraction of 13% (i.e. 864 patients sampled to the subcohort). By contrast the RS approach with 1239 patients results in a power of only about 40%. The strongly reduced power of the RS approach relates to the strongly decreased number of events included. Whereas for the NCC and CC design all individuals with an event are included, for the RS approach only individuals with the event of interest that are randomly sampled to the subcohort are included. For the RS with 1239 patients the mean number of events included is 82 (Table 1) out of 432 events.

5. Discussion

Although Figure 2 points out that the power of a NCC and a CC design is almost identical for an equal number of distinct persons involved we found a surprisingly clear benefit of the NCC design compared to the CC design in our data analysis. With
For the NCC design this means that individuals who are still at risk at later time points have a high probability of being sampled multiple times, therefore we have a high overlap of sampled controls and a reduced number of distinct persons. On the other side within the CC design the probability that an individual who is sampled randomly to the subcohort at the outset of the study is still at risk at a later time point is relatively small, thus little information might be contained within the subcohort regarding later event times.

Langholz and Thomas [27] compared the NCC and CC design with respect to the asymptotic relative efficiency. In accordance with our findings they state that the CC method can do “substantially worse” than the NCC method if there is moderate random censoring or staggered entry. Breslow et al. [28] derived a formula for the efficiency for β estimation in a NCC design relative to the full cohort. If a constant probability of exposure among controls is assumed regardless of the risk set (reasonable in the situation of a low event rate) it is straightforward to calculate this efficiency. Using the standard error of analysis of the full cohort in our ICU data set and the efficiency according to Breslow et al. [28] we calculated the predicted standard error for the NCC design and found that the results are in good agreement with the empirical standard errors we received in our NCC analysis (Table 1).

Langholz and Thomas [29] comment on sampling-induced covariance between score terms leading to a possibly smaller efficiency of the CC design compared to standard NCC sampling. Yet Borgan and Samuelsen [30] state that the efficiency of a NCC and a CC seem to be about the same if they involve the same number of distinct persons.

For a NCC design a fixed number of controls are sampled randomly from those at risk at each event time. Therefore a time scale (for example days since diagnosis or age) has to be fixed before sampling of controls and a later change in the time scale is not possible. In this regard the CC design offers more flexibility, as the subcohort (the source of controls) is sampled at the outset of the study without considering time or event status. Hence a benefit of a CC over a NCC is that multiple time scales are possible for analysis. One further consequence of the NCC sampling depending on the at risk status of individuals at each event time is that for every outcome a new sampling of controls is required. This is in contrast to CC sampling where sampling of the subcohort is done without respect to time or disease status and therefore it is possible to evaluate multiple disease outcomes using the same subcohort. This seems to be the key benefit of the CC over a NCC design. For example Kulathinal et al. [31] report on their practical experiences with respect to the case-cohort design. The authors point out that they opted for a case-cohort design.
for their genetic study because interest is in several outcome events, apart from death of any cause also non-fatal coronary heart disease and stroke events.

In order to overcome the aforementioned limitations of the NCC design there are several suggestions on methods where the sampled controls are no longer tied to their cases so that it is possible to analyze multiple endpoints using NCC sampling, see for example [32–34]. Yet note that the reuse of controls leads to a more complicated weighting scheme compared to standard NCC sampling.

The question concerning the choice of controls has also been the subject of several papers in this journal. For microarray studies Repsilber et al. [35] present a framework to deal with selection problems in microarray studies. They comment on the choice of microarray studies designed as case-control studies or as comparisons of parallel groups from cohort studies. Principles of comparability are highlighted,

![Figure 4](image-url) **Figure 4** Histograms of the parameter estimate (log hazard ratio) obtained from 100 bootstrap samples for the simple random sampling, CC, NCC design and the full cohort. The dashed vertical line yields the mean estimate obtained by analyzing the full cohort.
including the study-base principle in respect of homogeneity of samples and references. Stürmer et al. [36] describe flexible matching strategies to increase power and efficiency of case-control studies. Methods are presented to calculate power and relative efficiency compared to an unmatched design. However only static matching is considered, not dynamic matching like a nested case-control approach. Schröder et al. [37] comment on individual matching in observational studies. Here matching is used to select comparable subgroups on which further data analysis can be concentrated. Also in this case only static matching is considered. In contrast Wolkewitz et al. [38] examine dynamic matching with respect to a time-dependent exposure. Yet our approach focuses on dynamic matching with respect to an observed outcome.

6. Conclusion

We explicitly recommend using sampling designs like a nested case-control or case-cohort design not only for epidemiological but also for clinical cohort data, if a large cohort with a low event rate is present. In our hospital infection data example with only a small number of observed events and a high competing event rate, we showed that NCC and CC designs yield estimates comparable to estimates obtained by analysis of the full cohort while strongly decreasing the number of patients required. In contrast we clearly dissuade from using a simple random sampling approach that cannot be seen as an efficient approach for analyzing cohorts with only a small number of observed events.

For comprehensive comparisons of NCC and CC design we refer to [6, 27, 29]. Using the connection to the well-known Schoenfeld formula [4] the power of a NCC and CC design can be calculated straightforward in order to predict the loss of power coming along with a strongly decreased sample size compared to full cohort analysis.

With respect to the choice of the sampling design we sum up that the key advantage of the CC design is its high flexibility as multiple outcomes can be analyzed using the same subcohort. On the other hand the NCC design seems favorable in situations with a high censoring or competing event rate, becoming apparent in our ICU data analysis where for NCC sampling smaller standard errors and at the same time a lower number of distinct persons is achieved. In the end both the NCC as well as the CC design seem to have their relative merits and all in all provide a powerful and economical tool with respect to analysis of a large clinical cohort with only a few individuals failing of the event of interest.

Acknowledgments

Kristin Ohneberg was supported by Grant BE 4500/1-1 of the German Research Foundation (DFG).