Reuse of controls from nested case-control studies in cancer research

Nathalie C. Støer

Dissertation presented for the degree of
Philosophiae Doctor (PhD)

Department of Mathematics
University of Oslo
2013

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1459

ISSN 1501-7710

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika Publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

Acknowledgement

I formally started my work on this thesis in February 2011, however, informally already in August 2010. From August to February I was financed by The University of Oslo Graduate School in Biostatistics, interrupted by two months at the Department of Genes and Environment at the Norwegian Institute of Public Health, working with Hyperemesis Gravidarum. From February 2011 and throughout this project I have been financed by the The Norwegian Cancer Society. During these years I have been affiliated with the Division of Statistics and Insurance Mathematics/Statistics and Biostatistics at the University of Oslo. I have also spent at least one day a week at the Department of Chronic Diseases at the Norwegian Institute of Public Health.

My main supervisor has been Sven Ove Samuelsen with Haakon E. Meyer and Ørnulf Borgan as co-supervisors. First of all I want to thank Sven Ove who has followed me from my bachelor project during the third year of my studies, through the master thesis and up to now. He is also a co-author on all four papers comprising this thesis. I think it is safe to say that I would never have continued with a PhD if it had not been for you, as the idea had never crossed my mind before you brought it up. You are a sincerely kind person with a lot of (strange) knowledge. You answer all my questions, stupid and clever, with patients and your informal way of being and all digressions are much appreciated. I also want to thank my co-supervisor Haakon who is a co-author on Paper III, for his enthusiasm and good mood, and Ørnulf for his eye for details and vast experience. I must also thank Tom K. Grimsrud for all the effort he put in to providing me with the data for Paper IV and arranging for me to stay at the Cancer Registry of Norway during preparation of the data files. And finally Jan F. Bjørnstad for important knowledge on survey sampling, and Anders Skrondal for input on some of the papers, your ability to see details is impressive.

I would like to thank the other PhD-students at the 8th. floor for creating a social work environment. I am also grateful for the opportunity I have had to spend time at the Department of Chronic Diseases and be a part of the friendly and inspiring environment, and of course for the ice-cream breaks with Maria.

Finally, a big thanks to my friends and family. To my mum for always listening to my frustrations and to my friends for making me focus on "non-academic things", especially to Mathea for all her surprises and "great ideas" and of course to all my friends back home, you make vacations busy and fun.

Blindern, November 14th 2013
Nathalie C. Støer

List of papers

Paper I

Støer NC and Samuelsen SO (2012). Comparison of estimators in nested case-control designs with multiple outcomes. *Lifetime data analysis* **18**(3):261–283.

Paper II

Støer NC and Samuelsen SO (2013). Inverse probability weighting in nested case-control studies with additional matching - a simulation study. *Statistics in Medicine* **32**(30):5328–5339.

Paper III

Støer NC, Meyer HE and Samuelsen SO (2014). Reuse of controls in nested case-control studies. *Epidemiology* **25**(2):315–317.

Paper IV

Støer NC and Samuelsen SO. multipleNCC: Inverse probability weighting of nested case-control data in R. *Manuscript.*

Contents

1 Introduction

In epidemiology, cohort studies are often used to investigate relationships between exposure and outcome. The subjects in the cohort are followed from inclusion in study to time of event or censoring. Fairly inexpensive covariate information can often be obtained for all cohort members from registries containing standard measurements like for instance height, weight or BMI, or from questionnaires. However, in many situations such information is not sufficient to carry out a particular study and additional information must be collected. This can be expensive, time consuming and even logistically unfeasible for large cohorts. An alternative is to sample a subset of the cohort and collect covariate information only from the sampled subjects.

The classical case-control design (Breslow, 1996), which has been used by epidemiologists since the 19th century (Lilienfeld and Lilienfeld, 1979), is a popular and natural way to sample a subset from a larger population. Subjects experiencing the event of interest, referred to as cases, and controls, which are subjects not experiencing the event, are sampled separately, usually with a higher sampling fraction for cases than for controls. The main underlying idea is that a case is more informative than a control when the disease is rare. Therefore, including many or all of them, but only a fraction of the subjects not experiencing the event is a reasonable and efficient strategy. Sometimes the controls are sampled by simple random sampling, but often more sophisticated sampling designs which involve stratification or matching are adopted.

The standard unmatched case-control design does not incorporate time. Time is, however, often important since the incidence of many diseases increase with time (age). Two designs related to the standard case-control which incorporate the time aspect are the nested case-control (NCC) (Thomas, 1977) and case-cohort (CC) (Prentice, 1986; Kalbfleisch and Lawless, 1988) designs. With a nested case-control design, at each event time m subjects who have not yet experienced the event is sampled from the subjects at risk, and we say that the controls are matched on time or matched on at risk status. With a case-cohort design, a subcohort is sampled at the outset of the study and used as control group at all event times.

The nested case-control design has traditionally been more popular than the case-cohort design among epidemiologists. The reason for this might be that the NCC-design has been considered easier to analyze. However, the case-cohort design has the advantage that it is straightforward to use the same controls for a different endpoint, since the subcohort is a random sample from the cohort. In contrast, it has traditionally been considered impossible to reuse controls from a NCC-design since the cases and controls are matched on at risk status and potentially additional factors.

However, Samuelsen (1997) proposed a method for breaking the matching which allows NCC-data to be analyzed as CC-data. This method involves calculating the probability of ever being sampled as a control, and then weighting the controls with the inverse sampling probability, referred to as inverse probability weighting (IPW). This has later been discussed by a number of authors (Suissa et al., 1998; Kim and De Gruttola, 1999; Chen, 2001; Samuelsen et al., 2007; Saarela et al., 2008; Salim et al., 2009; Cai and Zheng, 2012; Chen et al., 2012; Salim et al., 2012; Pugh et al., Unpublished report).

Reuse of controls is useful when two or more endpoints are analyzed within the same or overlapping cohorts, because the controls for one endpoint can be used as controls for the other endpoint. Another situation where reuse of controls can increase efficiency is when analyses of subsets of original cases are carried out. Then, with the traditional method of Thomas (1977), only the controls sampled for the cases experiencing the sub-endpoint can be included in the analysis, ignoring all other sampled controls. Unlike with an IPW analysis, all sampled controls can be used as controls for the cases in question, and this can sometimes increase the number of controls drastically. Another advantage of the IPW analysis is that the time scale can be chosen after the controls have been sampled, while for the traditional method the time scale must be determined before sampling the controls, and cannot later be changed. Moreover, it allows for other models than proportional hazards models, and might be considered somewhat more general than the traditional method.

Alternatives to IPW for increasing the efficiency in NCC-designs have been suggested. Scheike and Juul (2004); Saarela et al. (2008) introduced full likelihood methods, where covariates only known for cases and controls were regarded as missing for non-sampled subjects. A somewhat similar idea was introduced by Keogh and White (2013), they however, handled the missingness by multiple imputation.

A problem with nested case-control designs in connection to IPW is additional matching. Often the controls are matched to the cases on additional factors, and this complicates the estimation of sampling probabilities. Additionally, the matching variables are usually confounders and breaking the matching will introduce confounding if the matching variables are not properly adjusted for. Some authors have used IPW in situations with additional matching (Salim et al., 2009; Cai and Zheng, 2012; Salim et al., 2012), however, they do not discuss the extra complexities this introduce.

A number of authors have used the NCC-design in a variety of fields. A few examples are Hultman et al. (1999) who studied the risk of schizophrenia and psychosis in relation to pregnancy and perinatal characteristics in a matched nested case-control study and Cnattingius et al. (1999) looked at preterm birth and the risk of anorexia nervosa. Floderus et al. (1993) and Tynes and Haldorsen (1997) studied exposure to electromagnetic fields and leukemia and brain tumors while Grimsrud et al. (2002) investigated the association between nickel exposure and risk of lung cancer. Other examples include Parsonnet et al. (1991); Øyen et al. (1997); Hankinson et al. (1998); Juul et al. (2002); Levine et al. (2004); Pischon et al. (2004); Dahm et al. (2010); Clendenen et al. (2011); Meyer et al. (2013). In all of these studies truncation time, event/censoring time and potential matching variables are known for all cohort members, and in 10 of the studies there were either multiple endpoints, or sub-analyses were performed on smaller parts of the sampled data.

The intention of this thesis was to explore methods for reusing controls with emphasize on IPW, and then mainly with regard to additional matching. When I started this work there were only one paper (Salim et al., 2009) mentioning IPW with additionally matched data, as far as I know.

Since additional matching will complicate the estimation of weights and the analyses itself, and since matching on additional factors is often performed, a more thorough evaluation of IPW and additional matching was needed. A secondary goal was to introduce IPW to epidemiologists. It has been over 15 years since the first paper (Samuelsen, 1997) on IPW for NCC was published, however, it does not seem that epidemiologists yet have realized the potential. A reason for why the epidemiologists have not yet picked up on it could be the lack of existing software to carry out such analyses, thus weight estimation must be carried out "manually" in each study. A step in the right direction could therefore be to develop software that more automatically estimates weights and carries out weighted Cox-regressions.

The outline of the thesis is as follows: Important epidemiological designs are discussed in Section 2, particularly the nested case-control design. I also mention additional matching and multiple outcomes. The aims of the thesis are given in Section 3, while inverse probability weighting, both in general and in connection to the NCC-design is discussed in Section 4, together with weight estimation and calibration of weights. Section 5 presents some promising alternatives to IPW. A summary of the papers is given in Section 6, while Section 7 contains a final discussion of our findings and indications of possibilities for further work.

2 Epidemiological designs

2.1 Cohort design and analysis

A cohort is a well-defined group of study subjects, for instance all individuals born in Norway from 1900 to 1999, or all subjects participating in a specific survey. A cohort study is often referred to as a prospective study and the conceptual framework is to follow the members of the cohort forward in time for occurrence of disease. Some cohort members will be exposed to the risk/protective factor of interest, while other will remain unexposed. Some exposures are time-invariant for instance sex or ethnicity, while other may change over time e.g. smoking habits or physical activity. At the end of follow-up, how exposure influenced time until disease can then be analyzed by contrasting the unexposed subjects with the exposed subjects.

In survival analysis, the time from start of study to the event of interest is in focus. Let $\tilde{t}_i$ be time to event for subject i, however, all subjects may not experience the event during follow-up, and are instead censored at c_i. The censoring time is usually either end of follow-up, which may or may not be the same for all subjects, or the time when the subject was lost to follow-up for other reasons, such as death, emigration or withdrawal from study. The observed follow-up time t_i for subject i is therefore the minimum of the censoring time and the event time. Often, the subjects are not followed from time zero, but from some later time v_i. This may for instance happen if age is used as time scale, then v_i will be the age of subject i when it was included in the study. This is referred to as left-truncation or delayed entry and v_i as the left-truncation time.

Since the time until disease is of interest in survival analysis, a natural quantity to model is the hazard rate $h(t)$, which is defined as the instantaneous probability of experiencing the event at time t given that the event has not happened up to that time. Different models for the hazard rate have been suggested (Marinussen and Scheike, 2006; Aalen et al., 2008), but the most famous is the proportional hazards model (Cox, 1972)

$$h_i(t|\mathbf{x}_i) = h_0(t)\exp(\beta'\mathbf{x}_i).$$

Here $h_0(t)$ is the baseline hazard, the hazard when all covariates equals zero, $\mathbf{x}_i = [x_{i1},\ldots,x_{ip}]$ is the vector of covariates and $\beta = [\beta_1,\ldots,\beta_p]$ are log-hazard ratios. In this model the baseline hazard is left unspecified, and it is therefore referred to as semi-parametric. It is, however, possible to assume a parametric specification of $h_0(t)$, for instance with an exponential or Weibull distribution for the survival times. A more flexible approach is to change to the cumulative log hazard-scale and model the baseline hazard with spline functions (Royston and Parmar, 2002). In all these models, the covariates act on the hazard in a multiplicative way, and the hazard rates are restricted to be proportional (except for the Royston-Parmar model). Aalen (1980) proposed a non-parametric additive model for the hazard rate

$$h_i(t|\mathbf{x}_i) = h_0(t) + \gamma_1(t)x_{i1}(t) + \ldots + \gamma_p(t)x_{ip}(t)$$

where the parameters γ are arbitrary functions of time. Other flexible models for the hazard have been discussed by for instance Marinussen and Scheike (2006).

Cox's proportional hazards model (Cox, 1972) will be assumed throughout the thesis. The estimation of regression coefficients is based on a partial likelihood

$$L(\beta) = \prod_{j\in\mathscr{D}} \frac{\exp(\beta'\mathbf{x}_j)}{\sum_{i\in\tilde{\mathscr{R}}_j}\exp(\beta'\mathbf{x}_i)}.$$

Here $\mathscr{D}$ is the collection of all cases and $\tilde{\mathscr{R}}_j$ is the risk set at time t_j, thus $\tilde{\mathscr{R}}_j$ includes all subjects at risk just before t_j. The partial likelihood has similar properties as an ordinary likelihood, thus ordinary maximum likelihood theory can be applied to obtain estimates and standard errors (Andersen and Gill, 1982).

2.2 Sampling designs

In situations with rare events, one may need a very large cohort to obtain a sufficient number of cases. Obtaining (additional) covariate information for all cohort members can therefore, as mentioned above, be impractical, and performing cohort studies may be too ambitious. Another situation where full cohort analyses are undesirable is in studies using biological material, for instance blood samples stored in biobanks. Although such material usually is collected for the entire cohort, it is often intended for use in a number of different studies and it may therefore be difficult to access it for all cohort members. A way around such problems is to base the analyses on only a subset of the cohort. This is often performed by including all cases and sampling a subset of non-cases who act as controls. With the traditional case-control design, the cases are identified at end of follow-up and one must therefore look back in time to collect exposure information, hence the case-control design is often referred to as retrospective. The case-cohort and nested case-control designs, see Borgan and Samuelsen (2013) for an overview, try to overcome this problem by collecting exposure information during follow-up.

2.2.1 Case-control designs

Breslow (1996) stated that

> *... my belief* [is] *that the contributions made by statisticians to the development of case-control methodology over the past 50 years have been among the most important of the many contributions they have made to public health and biomedicine.*

Thus the case-control design and successors have been, and still are, very important in epidemiology. Outcome dependent sampling has also been utilized in econometrics (Manski and Lerman, 1977; Cosslett, 1981) and there it is often referred to as choice-based sampling. The econometricians goal when utilizing such designs is often to investigate how some explanatory variables influence the probabilities of making specific choices, and for rare choices it may be simpler to sample choices than to sample decision makers, hence the name.

A case-control design is a general term for designs where cases and controls are sampled separately. However, it is also a more specific group of designs which do not take into account different length of follow-up. With this design, exposure values for a group of subjects having the disease in question, referred to as cases are compared to the exposure values of a group of subjects not having the disease, referred to as controls. The control group can be a completely random sample from the cohort, or it may be a m:1 matched design where each case have m controls matched on some background variables. It is not required that the cases and controls come from a fully defined cohort, however, they should be comparable on important factors and should also be representative of the population for which one wants to draw conclusions.

Due to the retrospective nature of starting with the effect (disease) and look for the cause (exposure), it was for a long time considered that the information from the cases and controls did not contain relevant information about the disease rate, which was the parameter of interest. However, Cornfield (1951) showed that the exposure odds ratio for cases and controls equals the disease odds ratio for exposed and unexposed. And further, when the disease in question is rare, the exposure odds ratio approximates the relative risk. Thus, a cohort analysis and the corresponding case-control analysis should yield approximately the same result when the disease is rare. Mantel and Haenszel (1959) put it down in words with the famous quote

> *A primary goal is to reach the same conclusions in a retrospective study as would have been obtained from a forward study, if one had been done.*

In the same paper a method for controlling for confounding in 2x2 tables, by stratifying into a number of sub-tables, was introduced.

It seems that the term "rate" was being used in a somewhat wider sense by for instance Cornfield (1951) than what is common today. He argues that the proportion of smoking subjects in the general population having cancer of the lung, is the proportion of smoking subjects among those with cancer of the lung in the sampled data multiplied by the proportion of the general population who have cancer of the lung during a specified time interval, and he states that this proportion is a rate. The time aspect is however only incorporated through the information from the general population and not from the sampled data. Thus, when considering the ratio of such proportions or rates, the contribution from the general population cancels and the time aspect is thereby no longer incorporated.

Today most case-control studies are analyzed with logistic regression and confounding is adjusted for by including the confounders in the regression model. This methodological development for unmatched studies was initiated in the paper by Cornfield et al. (1960). Further developments were provided by Cox (1966), Day and Kerridge (1967), Anderson (1972) until Prentice and Pyke (1979) finally demonstrated that the estimates from a logistic regression model for unmatched case-control data yields the desired estimates, and that the usual covariance matrix for logistic regression is valid.

For matched case-control studies Breslow et al. (1978), see also Prentice and Breslow (1978) and Hosmer and Lemeshow (1989, Chap. 7), developed a conditional logistic regression model which coincides with the traditional partial likelihood for NCC-data.

2.2.2 Nested case-control designs

In 1977, Thomas noted that most of the computational cost of carrying out a Cox-regression in a full cohort was connected to the censored subjects. To simplify the computational burden, he suggested taking a sample of each risk set. This has later become known as the nested case-control design. The important difference between this design and the traditional case-control design is that the controls are sampled from the risk set of each case. Thus the controls are required to be event free at the *time* the case experience the event and we say that the cases and controls are matched on time, or on at risk status. Usually the same number of controls, m, are sampled for each case, however, time-varying number of controls $m(t)$ are allowed. In addition to at risk status, the controls can be matched on other factors, for instance year of birth, sex or county of residence. The controls are sampled independently at each event time, thus a subject can be sampled as a control for multiple cases. A subject can also be sampled as a control and later itself become a case. These features are illustrated in Figure 2.1.

Thomas (1977) suggested maximizing a partial likelihood similar to the standard Cox-likelihood

$$L(\beta) = \prod_{j \in \mathscr{D}} \frac{\exp(\beta' \mathbf{x}_j)}{\sum_{i \in \mathscr{R}_j} \exp(\beta' \mathbf{x}_i)} \tag{2.1}$$

to obtain regression coefficients, and Oakes (1981) gave an argument for (2.1) being a partial likelihood. In (2.1) $\mathscr{R}_j$ is the sampled risk set at time t_j, hence the case at time t_j and its sampled controls. The vector $\mathbf{x}_i$ consists of the main exposure and adjustment variables. In practice, the estimation is carried out by a stratified Cox-regression, where the stratification is with respect to sampled risk sets. Maximizing (2.1) is what I will refer to as the traditional estimator for NCC-data. Note that time varying covariates are easily handled with this likelihood since the covariate values are only required to be known at the event time of the case in the risk set.

It took about 15 years from Thomas suggested the nested case-control design until the theoretical properties were fully understood. Goldstein and Langholz (1992); Borgan et al. (1995)

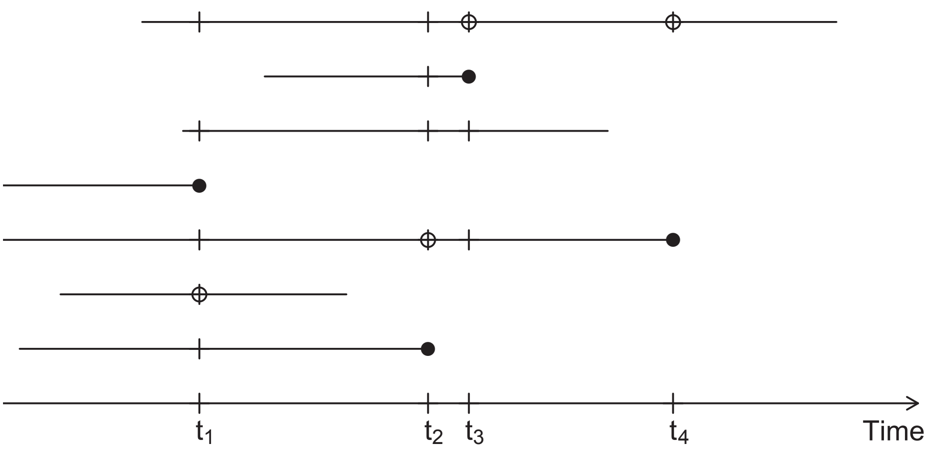

Figure 2.1: Example of nested case-control sampling with one control per case from a hypothetical cohort of 7 subjects. Each line represent the follow-up time for one subject. The beginning of the line is the left-truncation time and the end of the line is either the event time or the censoring time. Cases are represented by dots (●), possible controls by bars (|) and sampled controls with circles (○).

proved that inference could be based on large sample theory for the partial likelihood. Hence, the estimator is approximately normally distributed and the variance can be obtained from the inverse of the information matrix.

The efficiency of a nested case-control design is approximately $m/(m+1)$. When there is only one covariate and the regression coefficient equals zero, this holds exactly (Goldstein and Langholz, 1992). However, the efficiency may in general be lower and computations in Borgan and Olsen (1999) suggest that small fractions of exposed subjects and large relative risks decrease the efficiency.

In the traditional NCC-design all cases are included and the controls are sampled randomly with replacement from the (matched) risk sets of the cases. However, Borgan et al. (1995); Langholz and Goldstein (1996) have shown how more general sampling schemes can be analyzed by applying appropriate sampling weights, $w(t_j)$ to a partial likelihood on the form

$$L(\beta) = \prod_{j\in\mathscr{D}} \frac{\exp(\beta'\mathbf{x}_j)w_j(t_j)}{\sum_{i\in\mathscr{R}_j}\exp(\beta'\mathbf{x}_i)w_i(t_j)}. \tag{2.2}$$

An application of this is counter-matching (Langholz and Borgan, 1995), which is a stratified sampling design for NCC-data. The idea is to stratify the covariate of interest, or a surrogate variable, in say L strata and sample m_l controls from the n_l subjects at risk in strata l, except for the strata which the case belong to, where $m_l - 1$ controls are sampled. The m_l's may, but need not, be chosen proportional to the size of the strata. Sampling the controls in this way will (approximately) ensure a given covariate distribution within each sampled risk set, thus possibly increasing the variation of the exposure distribution, which may increase efficiency. Counter-matching uses all cohort information on the stratification variable, thus if the only covariate in the partial likelihood is the stratification variable, one can show that (2.2) is the corresponding cohort likelihood (Langholz and Borgan, 1995).

It is interesting to note that it is not clear who coined the term "nested case-control design". The term is sometimes used by epidemiologists when referring to any case-control design nested within a cohort, it has therefore proven difficult to establish the first use of the term in connection to Thomas' design. The first time the term appears in the Web of Science database is in an abstract to a conference for the Society for Epidemiological Research (Bond and Flores, 1984). However, they do not define the term in the abstract and from a personal communication with Bond it appears that the use of the term was in a broader sense. The first paper I could find that refers to Thomas (1977) and also uses term "nested case-control" is Lubin (1986). He, however, does not define the term either, which leads me to believe that design was not named by him. We tried asking Bryan Langholz, who did not know, and as a last resort I contacted Duncan Thomas, as I thought that he ought to know. From a person communication with him it appears that he, however, does not, and that he actually has been doing some digging himself to find it out.

2.2.3 Case-cohort designs

About 10 years after Thomas first suggested the nested case-control design, the case-cohort design was proposed (Prentice, 1986; Kalbfleisch and Lawless, 1988). It differs from the NCC-design in how the controls are sampled, instead of sampling from the risk sets, a subcohort is sampled at the outset of the study and used as control population at all event times. Covariate information is obtained for the subjects in the subcohort and for the cases occurring outside. Sampling a subcohort at the outset of the study is advantageous since it can be used as control population for different types of cases, and also because asserting covariate values can begin at start of follow-up.

Different estimation procedures have been put forward. Prentice (1986) suggested only including the cases outside the subcohort at their event time, thus maximizing a pseudo-likelihood on the form

$$L(\beta) = \prod_{j\in\mathscr{D}} \frac{\exp(\beta'\mathbf{x}_j)}{\sum_{i\in\tilde{\mathscr{S}}_j}\exp(\beta'\mathbf{x}_i)} \tag{2.3}$$

where $\tilde{\mathscr{S}}_j$ contains the subjects at risk in the subcohort at time t_j including the case at t_j if it did not belong to the subcohort. The inference for (2.3) was later justified by Self and Prentice (1988). Note that each term in the product in (2.3) is on the same form as the terms in (2.1), however since the same controls are used repeatedly, the terms are correlated (Langholz and Thomas, 1991) and (2.3) is therefore not a partial likelihood.

Another possibility is to include all cases whenever they are at risk. The estimation can then be based on a weighted partial likelihood (Kalbfleisch and Lawless, 1988)

$$L(\beta) = \prod_{j\in\mathscr{D}} \frac{\exp(\beta'\mathbf{x}_j)}{\sum_{i\in\mathscr{S}'_j}\exp(\beta'\mathbf{x}_i)w_i}. \tag{2.4}$$

Here the $\mathscr{S}_j'$ is the collection of subjects in the subcohort and all cases outside the subcohort at risk at time t_j. The weights are the inverse sampling probabilities $w_i = 1/p_i$ where p_i is the probability for subject i of being included in the subcohort. This probability is 1 for cases, when we choose to include all of them. For the non-failures Chen and Lo (1999) and Borgan et al. (2000) suggested, to use the fraction of non-failures in the subcohort compared to all non-failures in the cohort. See Cologne et al. (2012) for a review of existing estimation methods for the case-cohort design.

Since the subcohort is used repeatedly at each event time, the likelihood contributions are correlated and the inverse of the information matrix cannot be used to estimate the variance. A simple solution is to use robust variances (Lin and Ying, 1993; Barlow, 1994). An alternative is the plug-in estimators (Therneau and Li, 1999; Langholz and Jiao, 2007).

There are often some covariates which are known for all cohort members. If some of these fully observed variables are correlated with the main exposure, this can be utilized to increase the efficiency by stratified sampling (Borgan et al., 2000). The idea is to stratify the cohort in, say L strata, according to the fully observed variables and sample separate subcohorts from each stratum. The total subcohort is then constructed by combining the separate subcohorts, and this may increase the variation of the exposure distribution, which in turn may increase the efficiency compared to simple random sampling. The estimation can for instance be based on Estimator II in Borgan et al. (2000); Samuelsen et al. (2007) which uses (2.4) with modified weights. The weights for subjects belonging to stratum l are estimated as the number of non-failures in the stratum in the cohort divided by the number of sampled subjects in the given stratum. The strata are redefined so that the cases make up their own stratum with sampling probability 1. In this case the robust variance estimator can be conservative and a modified plug-in variance estimator (Langholz and Jiao, 2007; Samuelsen et al., 2007) should be used instead.

2.3 Additional matching

A common way of handling confounding is to adjust the estimate of interest for the confounders by including them in the regression model. Another approach is to only sample controls that have the same or similar value as the case on one or more variables. This is known as additional matching, and is often performed with NCC-data.

I will divide the matching criteria into two main methods; category matching and caliper matching (Cochran and Rubin, 1973). With category matching, the matching variable is usually categorical (with a fairly low number of levels) and the cases and controls must match exactly on the given variable. Examples of this can be sex or county of residence. With caliper matching, the control's matching variable must lie within a specified interval around the case's matching variable to be considered as a potential control. This type of matching criterion is often preferred when the matching variable is continuous or has a large number of ordered levels. Examples of such matching can be date of birth ± 12 months or month of blood sampling ±3 months.

With the traditional estimator for NCC-data, the sum in the denominator of (2.1) is over the sampled risk set, thus the subjects in each risk set will have equal, or similar values of the matching variables. With category matching, the matching variables do therefore not need to be included in the regression model since their contribution cancels in (2.1). The matching variables are often also ignored with caliper matching, hence the estimation is usually carried out as if no matching had been applied. However, disregarding the matching variables with caliper matching can yield biased estimates if the confounding effect is strong, and the matching intervals are too wide to fully capture the confounding.

2.4 Multiple outcomes

Sometimes more than one endpoint is of interest in a study. It may be planned on beforehand to study different endpoints, or retrospectively one wants to (re-)use the sampled data for other types of events or sub-endpoints. In situations with multiple endpoints and nested case-control data, being able to utilize all sampled controls through IPW may increase efficiency.

Different "types" of multiple outcomes occur in different situations. When two or more mutually exclusive events are of interest, the result is a competing risks situation, Figure 2.2. An example of this can be death from cancer and death from cardiovascular diseases.

In many situations, the main endpoint can be divided into meaningful sub-groups and separate analyses of each sub-endpoint can be of interest. It could for instance be informative to analyze metastatic and non-metastatic cancer separately, or divide cardiovascular disease into ischaemic heart disease and stroke. Such sub-endpoints can be seen as a special type of competing risks situations and utilizing all sampled controls when analyzing each endpoint is desirable.

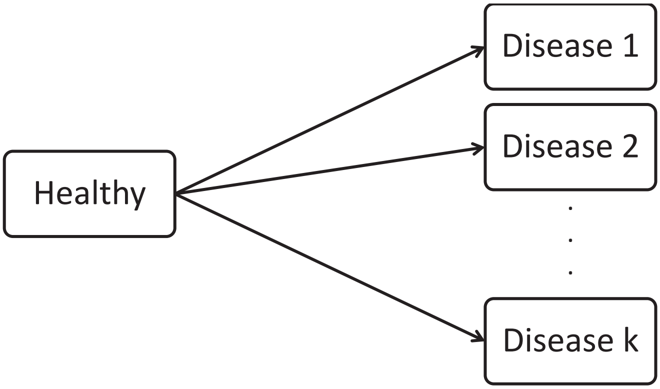

Figure 2.2: Competing risks.

A somewhat different multiple events situation I came across in Paper II and III is something we chose to call "subsequent events", Figure 2.3. In such situations a second, or subsequent, event can only happen after the first event has occurred. The example from Paper II and III is incidence and death from prostate cancer. Death from prostate cancer cannot occur unless the subjects developed prostate cancer at an earlier point in time. When considering the subsequent endpoint in Paper II and III, we analyzed the time from inclusion in study to time at death or censoring, and used the already sampled controls for the incident cases that also died from prostate cancer as controls. These controls are sampled at the time the cases experience prostate cancer, thus at an earlier point in time than the event of interest. This is not problematic with IPW analyses since the matching between cases and controls is broken, thus the time of sampling is irrelevant. We argued that it also will be valid for the traditional estimator as long as the controls do not change their behavior in any way after they have been sampled as controls. Another potential problem with the traditional estimator when using controls sampled before the cases experienced the event, is that some controls may already have been censored at the event time of their case. Those controls are of course excluded, and it can therefore be viewed as a situation with time dependent number of controls.

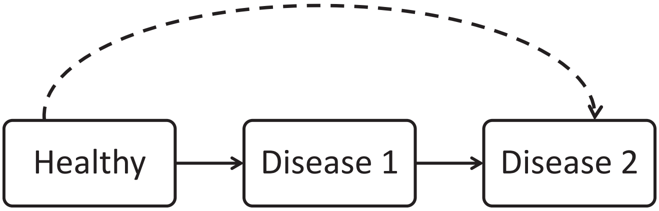

Figure 2.3: Subsequent events.

Another situation where IPW can be advantageous is when cases and controls are stratified with respect to a background covariate into sub-groups, and separate analyses within each stratum are performed. For instance in a previous analysis of the vitamin D and prostate cancer data (Meyer et al., 2013) it was considered reasonable to stratify the cases and controls into season of blood sampling (e.g. winter, spring, summer and autumn), since it is well known that sun exposure influences the vitamin D levels, and the potential for sun exposure varies with season in a country like Norway. Even though the controls were matched to the cases on date of blood sampling, the matching was not close enough to prevent some case-control pairs to be separated into different strata. With the traditional estimator, only the case-control pairs where both the case and the control fall into the same strata will contribute to the analysis. With IPW on the other hand, case-control pairs can be separated but still contribute in their respective strata.

The common feature of multiple endpoints situations is that the traditional estimator can only utilize controls matched to the cases in question. Therefore, controls with fully available covariate information may have to be disregarded. For instance in competing risks situations; controls for competing endpoints cannot be utilized. And in subsequent events situations, only the controls for the initial cases that also experience the subsequent event can be included. Excluding information will generally increase the uncertainty and it is therefore advantageous to be able to utilize all sampled controls in every analysis. The IPW method allow for breaking the matching and does therefore permit such use of controls, which in turn may increase the efficiency of the NCC-design.

I will in the following allow for more than one type of endpoint in the cohort by assuming a competing risks type of situation, and I will assume outcome specific proportional hazards models for each type of event $k, k = 1, \ldots, K$. The model for each endpoint is then on the same form as in a single event situation, i.e. the cause specific hazard for cause k is given by

$$h_{ik}(t|\mathbf{x}_i) = h_{0k}(t)\exp(\beta_k'\mathbf{x}_i)$$

where h_{0k} is the cause specific baseline hazard and $\beta_k = [\beta_{k1}, \ldots, \beta_{kp}]$ are the cause specific log-hazard ratios. The estimation of the regression coefficients connected to the $k-$th endpoint can therefore be based on maximizing

$$L_k(\beta_k) = \prod_{j \in \mathscr{D}_k} \frac{\exp(\beta_k' \mathbf{x}_j)}{\sum_{i \in \tilde{\mathscr{R}}_j} \exp(\beta_k' \mathbf{x}_i)}. \tag{2.5}$$

Here $\mathscr{D}_k$ consists of the cases of type k. The (2.5) is the likelihood for cohort data, but the same argument apply to NCC-data and the likelihood is given by substituting $\tilde{\mathscr{R}}$ with $\mathscr{R}$.

The competing risks and subsequent event situations are simple and transparent situations. The picture might however be much more complex, for instance with combinations of competing risks and subsequent events, Figure 2.4. In such situations, the simple approach by carrying out one Cox-regression for each endpoint is probably too simplistic for the problem at hand. However, using all sampled controls for the analyses may still be advantageous. An example of a situation as in Figure 2.4 can be overweight (Disease 1), cardiovascular diseases (Disease 2) and death. Overweight may lead to cardiovascular diseases which might lead to death. Subjects might however experience cardiovascular diseases without being obese and obesity might lead to death through other diseases than cardiovascular diseases.

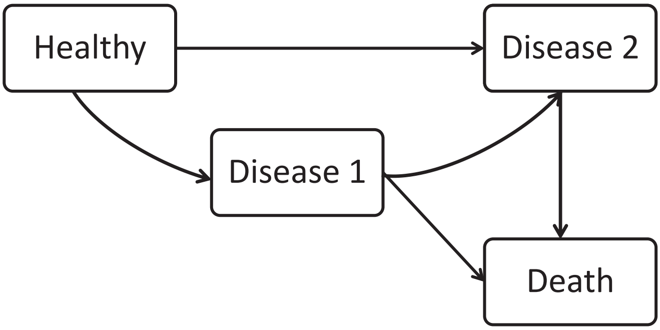

Figure 2.4: Simple example of a more complex situation.

3 Aims of the thesis

The overall topic of this thesis has been reuse of controls in nested case-control designs with a special focus on inverse probability weighting. The aim has been twofold; investigating the properties of the weighted partial likelihood estimator, especially in settings with additional matching, and try to make inverse probability weighting in nested case-control designs more accessible for epidemiologists.

For the first aim; the experience with IPW is yet fairly limited and gaining more insight through data examples is important. Multiple endpoints have been central since the potential for efficiency gain is larger in such situations. In the first paper we considered competing risks situations and compared IPW with other alternatives for reuse of controls, namely maximum likelihood and the survey sampling technique calibration. A main focus has, however, been on additional matching since NCC-studies often are matched on additional factors. Looking into how the weights should be estimated in situations with additional matching, how they are affected by incorporating this additional matching and how this influence the final hazard ratios and estimated standard errors have been central. Another important aspect has been how the matching variables should be handled in the Cox-regression. Since the initial matching is broken with IPW, somehow including the matching variables in the regression seems intuitively important. These issues have not previously been carefully discussed by authors performing IPW with matched NCC-data.

The second goal has been to make it more appealing for practitioners i.e. epidemiologists to utilize the advantages of IPW. A large part of all conducted NCC-studies are carried out by epidemiologists, thus increasing their awareness of an alternative method for analyzing such studies and demonstrating that sometimes large efficiency improvement can be obtained, seemed relevant. IPW analyses require some programming to obtain sampling probabilities and also some reorganizing of data, which might prevent epidemiologist from utilizing it. Therefore, having a more automatic estimation procedure is desirable, and developing a "simple-to-use" `R`-package with similar syntax as the standard `coxph`-function for analyzing proportional hazards models, has been a goal.

4 Methodology

4.1 Inverse probability weighting

Weighting with inverse sampling probabilities originate from survey sampling and was originally used to obtain totals, means or other quantities of interest in a population. Hansen and Hurwitz (1943) appear to be the first to formally consider this possibility for sampling with replacement with unequal probability, and Horvitz and Thompson (1952) later generalized the idea to sampling without replacement.

The idea behind inverse probability weighting is that each sampled subject should represent a number of non-sampled subjects in the population. If all subjects are equally likely to be sampled, each sampled subject should represent equally many in the population. However, if some "group" of subjects is less likely to be sampled i.e. some subjects are sampled with smaller probability than the rest, these subjects should represent more population subjects, since those subjects likely are under-represented in the sample.

Let $\mathscr{S}$ be the set of sampled subjects, which in a nested case-control design will be cases and sampled controls. And further let y_i be a variable only known for the subjects in $\mathscr{S}$, and O_i the sampling indicator. We are interested in an estimate of the population mean of y_i which can be obtained with the Horvitz-Thompson (H-T) estimator

$$\hat{y}_{\text{H-T}} = \frac{1}{N} \sum_{i \in \mathscr{S}} w_i y_i = \frac{1}{N} \sum_{i=1}^{N} O_i \frac{y_i}{p_i}$$

with w_i being the inverse sampling probability of subject i, and N the population size. The Horvitz-Thompson estimator is not connected to a specific sampling design, and is therefore very general. Regardless of how the subjects are sampled, as long as the sampling probabilities are known, the H-T estimator is unbiased since $E[O_i] = p_i$. It may, however, have a huge variance and is therefore not always a good estimator for a population mean or total (Basu, 1988, Chap. XII) i.e. Basu's elephants, which is an example of estimating the total weight of circus elephants illustrating the pitfalls of the H-T estimator. It is known among survey samplers that in order to use the H-T estimator the y_i/p_i should be approximately constant for all i, thus y and p should be positively correlated and y and w negatively correlated. In such situations, the variance of the H-T estimator does not tend to be too large. If, however, y and p are not positively correlated, a large y-value could be weighted with the inverse of a small p, and this one observation may dominate the estimated mean.

To account for the unequal sampling fractions among cases and the rest of the subcohort in case-cohort designs, Kalbfleisch and Lawless (1988) suggested to maximize a weighted partial likelihood (2.4). This can be seen as an application of the Horvitz-Thompson estimator where the likelihood contributions are weighted with inverse sampling probabilities.

In a nested case-control study the controls are matched to the cases on follow-up time and potentially additional factors. In addition, usually all cases, while only a fraction of the subjects not experiencing the event, are sampled. Due to this, the NCC-data are not a random sample from the cohort, hence breaking the matching is not trivial.

To account for the non-random sample, inverse probability weighting can be applied also in this situation. The idea is to weight the controls in such a way that the weighted data set represents the full cohort, and then carry out a cohort analysis on the weighted data. The weights, which are inverse sampling probabilities, must be estimated from the available data. They will depend on length of follow-up, since the longer a potential control has been included in the cohort, the more opportunities it has had to be sampled as a control. In additionally matched studies, also the matching variables could influence the sampling probabilities.

The estimation of hazard ratios is based on a weighted version of the traditional Cox-likelihood (Samuelsen, 1997)

$$L_k = \prod_{j \in \mathscr{D}_k} \frac{\exp(\beta_k' \mathbf{x}_j)}{\sum_{i \in \mathscr{S}_j} \exp(\beta_k' \mathbf{x}_i) w_i}. \tag{4.1}$$

Here $\mathscr{S}_j$ is the set of cases and controls at risk at time t_j. This likelihood is identical to (2.4), thus we analyze NCC-data "as if it were CC-data". The sampling procedure is, however, more complex in the NCC-design, resulting in more complicated weights in (4.1). Note that the vector of covariates $\mathbf{x}_i$ may now include additional matching variables z_i.

As in (2.4), the likelihood contributions are not independent and the inverse of the information matrix cannot be used as an estimate for the variance. Robust variances (Barlow, 1994; Lin and Wei, 1989) are the simple solution to the problem. In our experience this is usually a good estimator of the variance, however there exist situations where it can be conservative. In an example given by Samuelsen et al. (2007) where the censoring times were proportional to the covariate, the robust variance was highly conservative. It is however unlikely that such a covariate is the main exposure in a NCC study.

Samuelsen (1997) provided a variance estimator for the Kaplan-Meier(KM) weights described below. The proposed estimator is on the form $\Sigma^{-1}+\Sigma^{-1}\Delta\Sigma^{-1}$. Where the first part corresponds to the variance if we have had $\sum 1/p_i \approx N$ independent observations, and the second part is the excess variance due to the sampling. The Σ is the covariance matrix of the score function and can be estimated with the information matrix of the weighted partial likelihood. The Δ can be estimated by

$$\hat{\Delta} = \sum_i \frac{1-p_i}{p_i^2} U_i'(\hat{\beta}) U_i(\hat{\beta}) + \sum_i \sum_{j,j\neq i} \frac{p_{ij}-p_i p_j}{p_{ij} p_i p_j} U_i'(\hat{\beta}) U_j(\hat{\beta}) \tag{4.2}$$

where the sums run over all sampled non-cases. The p_i is the probability of ever being sampled estimated with the KM-estimator below and p_{ij} is the simultaneous probability of sampling both subject i and j. The $U_i(.)$ is the score contribution for individual i. Note that the individual score contributions are asymptotically equivalent with the W_i's considered by Samuelsen (1997). Note also that $\hat{\Delta}$ is on the same form as the variance estimator of the Horvitz-Thompson estimator (Horvitz and Thompson, 1952).

A difficulty of the variance estimation is calculating the simultaneous sampling probability of two subjects i and j. Even though the sampling is carried out with replacement at each event time, the sampling probability considered is the probability of ever being sampled and indicators of this is not independent (Samuelsen, 1997). Since sampling one subject will decrease the probability of sampling another subject, the sampling indicators are negatively correlated. Explicit formulas for the covariance between sampling indicators $\widehat{\text{Cov}}_{ij} = p_{ij} - p_i p_j$ when no additional matching is present can be found in Samuelsen (1997). In situations with additional matching $\widehat{\text{Cov}}_{ij}$ can be estimated using equation (3.1) in Samuelsen (1997), replacing the number at risk with the number at risk that also meet the matching criterion and taking the product only over the cases where both i and j can be sampled. More details are given in Paper IV.

4.2 Weight estimation

The weights in (4.1) must be estimated from the data and several estimation procedures have been mentioned. Samuelsen (1997) suggested a Kaplan-Meier like estimator, see also Suissa et al. (1998), which later was generalized to handle additional matching (Salim et al., 2009; Cai and Zheng, 2012)

$$p_i = 1 - \prod_{j\in\mathscr{D}_k, i\in\mathscr{P}_j} \left\{1 - \frac{m}{|\mathscr{P}_j|}\right\}. \tag{4.3}$$

Here $\mathscr{P}_j$ is defined as the set of eligible controls for the case at time t_j $\{i : v_i < t_j < t_i, z_i \in [z_j - \varepsilon, z_j + \varepsilon], i = 1, \ldots, N\}$ and $|\mathscr{P}_j|$ is the number of elements in $\mathscr{P}_j$. The ε is a vector where each element represent one matching criterion. With category matching, the corresponding element in ε will be zero. For caliper matching, the element in ε will correspond to the matching interval, e.g. if a control is matched on time of blood sampling ±2 months the element in ε will be 2. The interpretation of (4.3) is that $m/|\mathscr{P}_j|$ is the probability of being sampled as a control for the case at time t_j, and the product over all event times of 1 minus such probabilities is the probability of not being sampled at any event times, thus 1 minus this is the required probability of ever being sampled as a control. The KM-weights can fairly easily be generalized to accommodate more complex sampling designs, for instance did Zheng et al. (2013) recently generalized them to account for quota sampling and Samuelsen et al. (2007) have earlier generalized them to counter matching.

More model based approaches for estimating the sampling probabilities, discussed by Robins et al. (1994); Pugh et al. (Unpublished report); Mark and Katki (2006); Samuelsen et al. (2007); Saarela et al. (2008) are logistic regression, referred to as GLM, and generalized additive models (GAM) (Hastie and Tibshirani, 2009, Chap. 9)

$$p_i = E[O_i | t_i, v_i, z_i] = \frac{\exp(\xi + f(t_i, v_i, z_i))}{1 + \exp(\xi + f(t_i, v_i, z_i))}. \tag{4.4}$$

The O_i is the sampling indicator, i.e 1 for sampled controls and 0 for non-sampled subjects in the cohort. The cases are excluded from the estimation since they are sampled with probability 1 by design. I have used $f(t_i, v_i, z_i) = f_1(t_i) + f_2(v_i) + f_3(z_i)$ for simplicity, but interactions could also have been considered. The $f_1(t_i)$ and $f_2(v_i)$ are smooth functions with GAM-weights and linear functions with GLM-weights. The $f_3(z_i)$ may take different forms, both with GAM- and GLM-weights, depending on the matching criterion. For category matching with not too many levels, the intuitive approach would be to include the matching variable as a categorical covariate. It is less intuitive how the matching should be handled with caliper matching because there is no explicit grouping since the intervals are formed around each case. One approach is to create categories, for instance of the same length as the intervals, and another approach is to include the matching variable as a continuous covariate in the logistic regression model.

A third weight estimator which I think is not well suited for additional matching is local averaging (Chen, 2001). Without additional matching this method involves choosing a partition of the time axis both with regard to inclusion time and censoring time. Let $0 = v^0 < v^1 < \ldots < v^A$ be a partition of the range of the left-truncation time and $0 = t^0 < t^1 < \ldots < t^B$ a partition of the range of the follow-up time where t_A and t_B is the upper limit of the left-truncation times and censoring times respectively. If we further define $\mathscr{J}_a = (v^{a-1}, v^a]$ and $\mathscr{I}_b = (t^{b-1}, t^b]$, the local averaging weights can be expressed as

$$w_{ab} = \frac{\sum_{i=1}^{N} I(v_i \in \mathscr{J}_a, t_i \in \mathscr{I}_b, i \in \mathscr{C} \backslash \mathscr{D})}{\sum_{i=1}^{N} I(v_i \in \mathscr{J}_a, t_i \in \mathscr{I}_b, i \in \mathscr{S} \backslash \mathscr{D})}. \tag{4.5}$$

when $\mathscr{C}$ denotes the collection of all cohort members. All controls included in the study in interval a with a censoring time in interval b are given weight w_{ab}. Hence, all subjects sampled within the same combination of intervals will be given the same weight. This corresponds to approximating the NCC-sampling with stratified random sampling where the strata are defined by inclusion times and censoring times, and Samuelsen et al. (2007) noted that it corresponds to post-stratification on follow-up time. Chen did not consider left-truncation, thus (4.5) is a slight generalization of the weight estimator he presented (Chen, 2001). Generalizing these weights to additional matching would require a partition of the matching variables in addition to the range of the left-truncation time and censoring time. This will easily generate a vast number of intervals with few subjects in each interval, thus the weights may become unstable.

Figure 4.1 displays a typical picture of the sampling probabilities for a situation without left-truncation and additional matching. The four estimators follow each other fairly closely. The KM- and GAM-probabilities cannot decrease with time, while the GAM- and Chen-probabilities can vary more freely and can both increase and decrease with time, which can be seen from the figure.

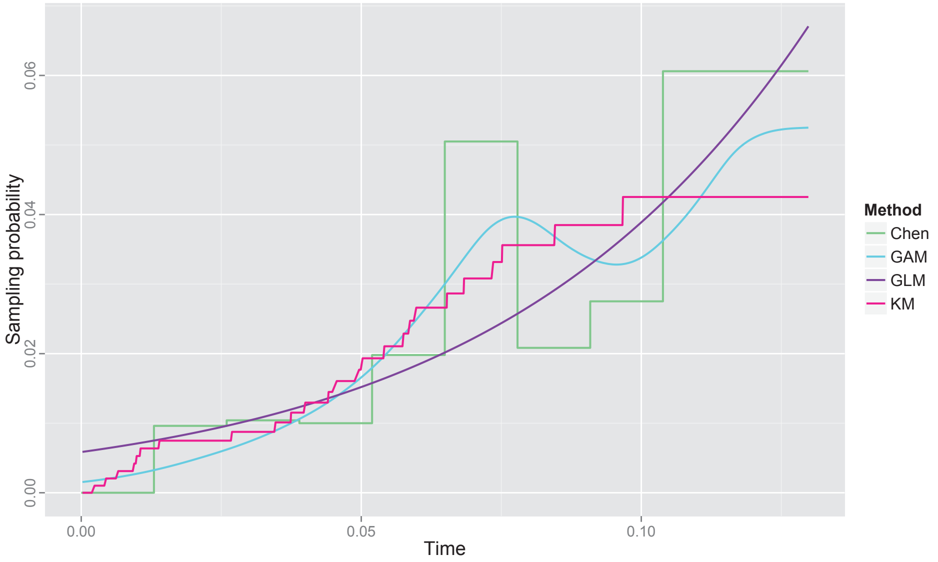

Figure 4.1: Estimated sampling probabilities as a function of follow-up time in a situation with no matching or left-truncation.

4.3 Calibration

Calibration originates from survey sampling and was first proposed by Deville and Särndal (1992) as a technique to improve the Horvitz-Thompson estimator for a population total $y_{\text{tot}} = \sum_{i=1}^{N} y_i$.

In many situations there exists so called auxiliary variables A_i, which are fully observed variables in the cohort correlated with the variable of interest y_i which is only known for the sampled subjects. In such situations, taking the weighted mean of the sampled subjects might not be efficient, since available information regarding the variables of interest is discarded. One way of utilizing the additional information is to incorporate A_i into the weights. This can be accomplished with calibration, by finding weights that fulfills the calibration equation

$$\hat{A}_{\text{tot}} = \sum_{\mathscr{S}} d_i A_i = \sum_{i=1}^{N} A_i = A_{\text{tot}}. \tag{4.6}$$

The d_i's are calibrated weights and have the property that the total of the auxiliary variable is estimated exactly. The idea is then that since A and y are correlated, $\hat{y}_{\text{tot}} = \sum_{\mathscr{S}} d_i y_i$ is probably closer to y_{tot} than $\hat{y}_{\text{H-T}}$.

The specification of the calibrated weights from (4.6) is not unique and an additional requirement is that the calibrated weights should be as close as possible to the original or crude weights. This requires a distance measure, $G(w,d)$ and Breslow et al. (2009b) suggested two alternatives

$$G_1(d,w) = \frac{(d-w)^2}{2w} \qquad \text{and} \qquad G_2(d,w) = d\log(\frac{d}{w}) - d + w.$$

Other distance measures are discussed in Deville and Särndal (1992) and Deville et al. (1993). The calibrated weights are thus weights that solves the calibration equation while minimizing $\sum G.(d_i, w_i)$.

Breslow et al. (2009a,b) suggested calibration as a method for reducing the variability due to the sampling in CC-designs. See also Lumley (2010) and Lumley et al. (2011). Ganna et al. (2012) used the calibration technique in a stratified CC-design and we informally generalized the idea to NCC-designs in Paper I.

The suggestion for auxiliary variables in Breslow et al. (2009a,b) is dfbetas $A_i = \Sigma^{-1}(\tilde{\beta})U_i(\tilde{\beta})$ where $\tilde{\beta}$ is the cohort estimate, $\Sigma^{-1}(\beta)$ the inverse of the information matrix and $U_i(\beta)$ the individual score contributions at the parameter value β. The reason for this choice of auxiliary variables is as follows: In CC- (or NCC-)analyses, the goal is not to estimate a population total with a weighted sum, but rather regression coefficients. Hence, using fully observed variables correlated with the exposure of interest is not necessarily a good idea. However, with a first order Taylor approximation of U_i around the true value of β we have

$$\hat{\beta} \approx \beta_0 + \sum_{\mathscr{S}} w_i \Sigma^{-1}(\beta_0) U_i(\beta_0)$$

both for CC- and NCC-designs when analyzed with a weighted partial likelihood. This expression is on the same form as the Horvitz-Thompson estimator, thus a good choice of auxiliary variables is something correlated with $\Sigma^{-1}(\beta_0)U_i(\beta_0)$, and the natural choice is cohort dfbetas.

Since full cohort information is not available, the cohort dfbetas must be estimated from the sampled data. By applying an estimation method suggested by Kulich and Lin (2004), the entire calibration analysis for (N)CC-data is a 5 step procedure. The practical details can be found in Breslow et al. (2009a) and in Paper I. The important part is that the partially observed variables are imputed using a regression model with fully observed variables as covariates, in order to estimate the cohort dfbetas. The stronger the association between the fully and partially observed variables is, the better the imputation is, and the more efficient the calibration technique become.

4.4 Simulation of survival data

In Paper I and II we conducted simulation studies. The main advantage of simulations contra analysis of real data is that important aspects of the data are under control when it is simulated from a given model. Hence, the true associations, and therefore also the parameters of interest are fully known. This makes it possible to quantify bias by comparing the mean of the parameters to the true value. Moreover, the variance estimator can be evaluated by comparing the empirical variance of the parameter estimates to the mean of the estimated variances. Another aspect often reported from simulation experiments is the coverage; the proportion of times the true value is contained in the confidence interval. A drawback with simulations is that it is hard to create simulation models that incorporate the complexities of data from the real world, thus simulations often create too simplistic situations with "well-behaved" data.

The first step when simulating survival times is to decide upon a parametric specification of the hazard function and combine this with a relative risk function. This must then in turn be translated into an expression for the event times. It is well known that a constant baseline hazard yields exponentially distributed survival times, while a monotonic increasing or decreasing baseline hazard function can correspond to Weibull distributed survival times. Bender et al. (2005) provide formulas for simulating from the exponential, Weibull and Gompertz distributions. Perhaps due to an easy relationship between the baseline hazard and survival times, the exponential or Weibull distribution is often used in practice. However, these distributions can be too simplistic and Crowther and Lambert (2013) describe a simulation technique where the baseline hazards can take almost any form.

The second ingredient is the censoring distribution. Since a survival time is the minimum of an event time and a censoring time, $t = \min(\tilde{t}, c)$, both an event time and a censoring time must be simulated for each subject in the cohort. The smallest value is defined as the follow-up time and a subject is considered to be a case if the event time was smaller than the censoring time, hence experienced the event during follow-up. The censoring distribution is often simpler than the distribution of event times, i.e. a constant, or uniformly distributed.

In Paper I, we simulated survival times from the simple exponential distribution with a uniform censoring distribution. While in Paper II, where we wanted to mimic the prostate cancer data, the Weibull distribution was considered adequate for the survival times. The censoring times in Paper II were the minimum of age at a certain date, representing end of follow-up and a Gompertz distributed variable representing the background mortality in the population.

5 Alternatives to inverse probability weighting

IPW is an intuitive and simple method for reusing controls, however, there exist other possibilities. These alternatives are based on taking on a missing data point of view. The full cohort is used in the estimation and the covariates only known for cases and controls are considered missing for non-sampled subjects.

5.1 Maximum likelihood methods

The maximum likelihood method of Scheike and Juul (2004) and Saarela et al. (2008) is based on the likelihood for full cohort data and covariates collected only for sampled subjects X are considered missing for the non-sampled subjects in the cohort. The full likelihood with no left-truncation can be expressed as

$$\begin{aligned} L(\theta,\mu) \propto & \prod_{i\in\mathscr{S}} p(T_i,D_i|X_i,z_i;\theta)p(X_i|z_i;\mu) \\ & \times \prod_{i\in\mathscr{C}\setminus\mathscr{S}} \int p(T_i,D_i|x,z_i;\theta)p(x|z_i;\mu)dx. \end{aligned}$$

Here $p(T_i,D_i|X_i,z_i;\theta)$ is the distribution of (T_i,D_i) conditional on (X_i,z_i) while $p(X_i|z_i;\mu)$ is the conditional distribution of X_i given z_i. The integral in the last product is due to the fact that X_i is unobserved for non-sampled subjects. Generalization to left-truncation is considered in Saarela et al. (2008). Note that z is now all fully observed variables included in the model and not necessarily matching variables.

Saarela et al. (2008) and Scheike and Juul (2004) differ in how they handle both $p(T_i,D_i|X_i,z_i;\theta)$ and $p(X_i|z_i;\mu)$. Saarela et al. (2008) take a fully parametric approach, thus assume a parametric specification of baseline, and specify the distribution of the partially observed covariates given the fully observed variables parametrically. Due to this parametric approach they are able to maximize the likelihood directly through numerical maximization. Scheike and Juul (2004) choose to keep the baseline non-parametric, as in a Cox-model, and do not require a parametric specification of the conditional distribution of the partially observed covariates. Thus this likelihood rest on fewer assumptions, but is harder to optimize than the likelihood of Saarela et al. (2008), and they therefor suggest to use the Expectation-Maximization procedure to obtain the estimates of interest.

For covariates only observed for cases and controls and when these covariates are (almost) uncorrelated with fully observed variables, the efficiency of Saarela's likelihood are often only marginally higher than the efficiency of IPW (Saarela et al., 2008). However, for fully observed variables or in situations where the partially observed covariates are correlated with fully observed variables, the efficiency of the full likelihood can be considerably higher.

Simulations in Scheike and Juul (2004) indicate that their likelihood become more efficient than the traditional estimator for increasing hazard ratios. This may be due to that the traditional estimator may lose efficiency for increasing hazard ratios (Borgan and Olsen, 1999) while the efficiency for the Scheike-likelihood seems to remain fairly constant. Another reason for the efficiency gain can be their large number of cases, which results in fully available information for a large portion of the cohort. Such efficiency gain is likely to be observed for the IPW approach as well, and in my master thesis (Støer, 2010, p.42-44) I showed that IPW has almost the same efficiency as the Scheike-likelihood when using their simulation setup. More formal comparisons between that likelihood and IPW has, however, not been conducted as far as I know, neither has comparisons between Saarela's likelihood and Scheike's likelihood.

Maximum likelihood methods for the proportional hazards model with other types of missing patterns has been considered by Chen and Little (1999) by a non-parametric estimator. With the aim of estimating haplotype-disease associations Zeng et al. (2006) proposed a likelihood method for NCC-data and Saarela and Kulathinal (2007) for case-cohort data. Scheike and Martinussen (2004) constructed a more general likelihood for case-cohort data in a similar manner as in Scheike and Juul (2004) for NCC-data, and Kulathinal and Arjas (2006) did similar things, only with a Bayesian perspective. It is worth noting that the full likelihood for NCC- and CC-data are identical, thus the sampling design is irrelevant after the data has been sampled when a full likelihood approach is chosen for analysis.

5.2 Multiple imputation

A somewhat different approach is based on multiple imputation (MI) (Marti and Chavance, 2011; Keogh and White, 2013). The general idea of MI originates from Rubin (1987) and the concept is that the missing covariates are drawn from some distribution conditioned on what is fully known. The analysis is then carried out on the full data set with the missing values imputed. To account for the extra uncertainty in imputing the missing values, the imputation is carried out a number of times, and the analyses are performed on each imputed data set. The individual point estimates are then combined into one final estimate, and the variance estimator takes into account the variability in predicting the covariates.

Marti and Chavance (2011) suggested using multiple imputation for case-cohort designs with a simple imputation model that did not account for time. Keogh and White (2013) extended the work to also include the nested case-control design and suggested more complicated imputation models that incorporate time. For the imputation, X is sampled from the conditional distribution of $X|z,D,T$. This distribution is non-standard and two approaches are suggested; approximate imputation models and rejection sampling.

From simulations in Keogh and White (2013): With one control per case the MI models have higher efficiency than the traditional estimator and the efficiency increases as the correlation between the covariate of interest and the surrogate measure increases. With 5 controls per case, the MI model is only marginally better than the traditional estimator. This is, however, as expected since the efficiency of the traditional estimator with 5 controls is high, thus the possible improvement is limited. Comparisons between MI and IPW have not been carried out as far as I know.

6 Summary of papers

6.1 Paper I

Støer NC and Samuelsen SO (2012). Comparison of estimators in nested case-control designs with multiple outcomes. *Lifetime data analysis* **18**(3):261–283.

The first paper compared IPW with four weight estimators, the maximum likelihood estimator of Saarela et al. (2008) and calibration through a series of simulations. Calibration had not previously been considered for the NCC-design, thus we also explained how the concept may carry over from the case-cohort design. Since the MLE approach relies on more modeling assumptions than IPW, we performed some simulations with misspecified models to check the sensitivity to the model assumptions. Finally, we also suggested a method for more efficient calculation of the likelihood of Saarela *et al.* by aggregating equal or similar covariate values and survival times. Additional matching was not considered in this paper.

The results from the comparison indicated that the MLE-method is more efficient than IPW for fully observed variables, and when the partially observed covariate was correlated with the fully observed covariate the efficiency was also improved for the partially observed covariate. The calibration approach was also more efficient than the IPW for the fully observed covariate, however, the achieved efficiency gains were more moderate than for MLE. For the partially observed variable, calibration gave only higher efficiency than IPW for the strongest association between partially and fully observed covariates. Estimates of interest were similar for all four weighting technique, MLE and calibration.

We further investigated the assumptions in the likelihood of Saarela *et al.* by misspecifying the conditional distribution of the partially observed covariate given the fully observed covariate, and erroneously specified the baseline hazard. In addition we also looked at a misspecification of the linear expression of the covariates. The results were that the MLE approach seemed to be fairly robust against misspecified baseline hazards; however, the modeling of the partially observed covariate given fully observed variables was important. For the last type of misspecification, the MLE approach fared no worse than any of the other methods.

Finally, we also found that the aggregation approach can decrease the computation time substantially in situations where the number of covariates is not too large, without introducing bias.

6.2 Paper II

Støer NC and Samuelsen SO (2013). Inverse probability weighting in nested case-control studies with additional matching - a simulation study. *Statistics in Medicine* **32**(30):5328–5339.

It is common to match the controls on additional factors. Although IPW has been applied to additionally matched data in a few papers (Salim et al., 2009, 2012; Cai and Zheng, 2012), it has not previously been carefully investigated how additional matching affects IPW estimation, moreover how the matching variables should be handled in the Cox-regression has not earlier been discussed. The aim in Paper II was therefore to consider some situations where we believed that additional matching could be problematic.

We conducted a series of simulation experiments where the fundamental setting was mimicking the vitamin D and prostate cancer data in Paper III. The controls were matched on date of blood sampling ± 2 months and age at blood sampling ± 6 months. We considered three potentially problematic issues; association between matching variables and exposure/outcome, close matching and batch effects, in addition to looking at a situation without any of the aforementioned issues.

When the exposure of interest is measured in batches, some variation between different batches due to the measurement equipment may occur. In such situations, measurements from the same batch will be more similar than measurements from different batches, and this is what we refer to as batch effects. For example, in the prostate cancer and vitamin D study the blood samples were analyzed in batches of 50 and the batches explained about 10% of the variation in the vitamin D measurements.

The results from the simulations were as follows: When there are associations between matching variables and exposure/outcome, the matching variables should be adjusted for in the Cox-regression. Including the matching variables in weight estimation was, however, not important in our simulations. Close matching seemed to be a smaller problem than we believed on beforehand. With GAM/GLM-weights we did not see any indication of bias even with extremely close matching, while with KM-weights there was some indication of bias in the most extreme situation. Batch effects, on the other hand, may attenuate the covariate effect when the matching is broken. We believe that the remedy for this problem could be measurement error methodology, however we did not pursue this in the paper.

6.3 Paper III

Støer NC, Meyer HE and Samuelsen SO (2014). Reuse of controls in nested case-control studies. *Epidemiology* **25**(2):315–317.

The third paper is a Research Letter including a more detailed web-appendix. It is an application of IPW in a situation with additional matching. It is also an attempt to reach the epidemiologists by explaining the concept of inverse probability weighting and weight estimation, and at the same time emphasizing that fairly large efficiency gains can be obtained by reusing controls.

We applied IPW in an additionally matched NCC-study (Meyer et al., 2013) of the association between vitamin D and incidence and death from prostate cancer. We found that IPW was more efficient than the traditional estimator, in particular for the death endpoint, and that the estimates were comparable to the estimates from the traditional estimator. We were also able to compare the traditional estimator and IPW with a cohort analysis by analyzing a covariate observed for all cohort members. The result was that the IPW estimates were closer to the cohort estimates than the estimates from the traditional estimator.

The web-appendix contains more details about the data and weight estimation. It also contains an additional analysis considering metastasis status. The incident cases were divided into three metastasis groups: Advanced cancer, localized cancer and unknown metastasis status. This can be seen as a competing risks situation since the groups are mutually exclusive. The discrepancy between the traditional estimator and the IPW estimators were similar to previous analysis, and the efficiency was close to or above 2 for all IPW estimators for all three endpoints.

6.4 Paper IV

Støer NC and Samuelsen SO. multipleNCC: Inverse probability weighting of nested case-control data in R. *Manuscript*.

One reason why IPW has not yet become a common tool in epidemiological research could be that software for carrying out the analyses does not exist. I have therefore constructed an `R`-package for calculating the weights and performing weighted Cox-regressions. The program also includes some possibilities for variance estimation. The entire estimation procedure can be performed with a one-line call with similar syntax as the `coxph`-function for analyzing proportional hazards models.

Paper IV is a demonstration of the `R`-package. We explain the estimation details and illustrate the use by analyzing a real NCC-study (Grimsrud et al., 2002) of nickel exposure and lung cancer and cancer of the nose and nasal sinuses. We also describe variance estimation with KM-weights with additional matching in some detail.

The estimated hazard ratios from the nickel data are fairly large, above 8 when comparing the second most exposed group to the least exposed group. Thus the paper offers some additional information regarding how IPW behaves in situations with large hazard ratios. Overall, large hazard ratios do not seem to be problematic. There are, however, some discrepancies between the estimates from the traditional estimator and the IPW estimates, but in relation to the standard errors, the discrepancies are not larger than expected. It could be anticipated that the efficiency of IPW would increase when the hazard ratios were large, since the efficiency of the traditional estimator may be lower than $m/(m+1)$ in such situations (Borgan and Olsen, 1999). The efficiencies were, however, similar to what we have seen in other analyses.

7 Discussion

The nested case-control design is a popular study design with over 5000 hits in the ISI Web of Science. Some of these hits may be due epidemiologists using the term more generally for any case-control study nested within a well-defined cohort. Nevertheless, it is a popular design among practitioners. The main advantage of NCC compared to a cohort analysis is the cost effectiveness. It is, however, inevitable less efficient than a cohort analysis and it is therefore important to use the available information as efficiently as possible. The main topic of this thesis has been reuse of controls by inverse probability weighting of the partial likelihood, which is one way to better utilize the available information. The main focus has been on additionally matched NCC-data, but I have also evaluated the different weight estimators, calibration and maximum likelihood in a situation without additional matching, and developed an `R`-package for performing IPW-estimation.

The contribution of Paper I is threefold; comparison between existing weight estimators for IPW analysis, indicate how calibration can be carried out with NCC-data, and study the MLE approach of Saarela et al. (2008) including an aggregation technique which may decrease the computational burden. For the first contribution, although similar comparisons between weight estimators already had been carried out by Samuelsen et al. (2007) and Saarela et al. (2008) with similar results, confirming once more that the weight estimators seem to be similar, and indicate that this also holds for left-truncation is useful. It is reassuring for practitioners that the estimates and variances are fairly consistent across weighting methods.

With regards to calibration of weights, which had not previously been considered for NCC-data, it may further improve the efficiency of IPW analyses. It is, however, important to note that the asymptotic properties of calibration with NCC-data have not yet been investigated. In Paper I, we only conjecture that the calibration properties of case-cohort data carry over to the NCC-data. In order to increase efficiency with calibration, fully observed variables correlated with the variables of interest must exist in the cohort. In situations with additional matching this is satisfied, and the matching variables will usually be correlated with both exposure and outcome. Investigating the calibration technique in such situations would be interesting, especially, since it may seem that IPW is somewhat less efficient in situations with additional matching.

For the last contribution from Paper I, we showed that the likelihood of Saarela *et al.* can be almost as efficient as cohort estimation when fully observed covariates correlated with variables of interest exist in the cohort. The experience with this likelihood is as far as I know fairly limited. We did, however, show that model misspecification can result in severely biased estimates. This is of course a drawback since the model assumptions are hard to validate. The likelihood can also be numerically demanding to evaluate, but our aggregation technique may decrease the computation time drastically in some situations.

The focus of Paper II is on additional matching. It was not obvious that IPW would continue to work properly in such situations. Since the probability of being sampled for the sampled controls increase towards 1 when the matching criterion gets narrower, the weights will decrease towards 1. With small weights, the idea of reconstructing the cohort by up-weighting the controls break down. We did, however, show through simulations that extremely narrow matching must be present before the KM-weights cause biased estimates and the GLM/GAM-weights are not affected by the matching criteria. Another side of this is that when we analyze NCC-data with IPW, we break the matching between the cases and controls. Our simulations showed that it is therefore important to adjust for the matching variables when they are confounders. However, matching on confounders and adjusting for them are two different things, although with the same ultimate goal, and again it was not trivial that IPW would continue to work when the matching was broken.

A problem with IPW, that will become more and more relevant as the number of data sets containing biological material for instance gene expression data or DNA methylation increases, is batch effects. This is an intrinsic problem of IPW and occurs because the accuracy of the analyzing equipment may vary and samples analyzed within the same batch will often be more similar than samples from different batches. The differences due to the batches do not pose a problem for the traditional estimator as long as the cases and controls belong to the same batch. However, when the matching is broken, as with IPW, we have seen from simulations that this may lead to an attenuating effect of the regression coefficients. Solutions to this problem can be adjusting for the batches or stratify according to them. Those solutions are however not satisfactory as the number of bathes can be large. Especially, stratifying on batches may reduce much of the gained efficiency with IPW. Another approach could be more traditional measurement error methodology as the addition or reduction due to the batches can be seen as measurement error. I have not tried this in practice so I do not know whether it would be a straightforward generalization or a more methodological challenging problem.

Armitage (1975) and Breslow and Day (1980) noted that unmatched analyses of matched data might cause conservative estimates. Thus, it is not new knowledge that analyzing matched data as an un-matched study may cause biased estimates. Our results regarding the importance of adjusting for the matching variables when the matching is broken is therefore not surprising. However, one could speculate that when the matching variables were included in the weights, the need for adjusting for the matching variables would vanish. This was not the case in our simulations. The reason for this is probably that including the matching variables in weight estimation only corrects for the biased sample with regards to the matching variables, and the confounding will still be present if the matching variables are not properly adjusted for.

In the simulations, IPW has generally been more efficient than the traditional estimator, even when the same number of controls were included in the estimation. The exception is in situations with additional matching when the same number of controls is used by both estimators. A possible explanation for the efficiency loss in such situations is that the efficiency gain due to the matching is lost when the matching is broken. Another reason might be that since the IPW-estimates must be adjusted for the matching variables, the IPW-likelihood contains more covariates than the traditional likelihood. Unless the extra covariates are independent of the covariates included in both estimators, the uncertainty of the IPW-estimates will increase.

In simulations without any additional matching in Paper I and in Samuelsen et al. (2007); Saarela et al. (2008), the KM-type of weights, GAM/GLM and local averaging produced only minor differences in final estimates and standard errors. While with additional matching, we have experienced somewhat larger differences between the GAM/GLM-weights and KM-weights The probable cause of this difference is that the additional matching is handled fundamentally different with the two types of weights. With GAM/GLM-weights, the matching variables are included as covariates, while for KM-weights, the matching criterion is used more explicitly when calculating the weights, and the KM-weights therefore reflect more closely the "true" sampling probabilities. However, with close matching the weights become small, perhaps even 1, and the idea of reconstructing the cohort breaks down. Then a more model based approach with GLM/GAM-weights might be better, since the weights do not become as small.

The weights used with additionally matched data are modified KM-weights and GAM/GLM with adjustment for the matching variables. Another possibility for the logistic regression type of weights is to do one regression per control with only left truncation times and follow-up times as covariates. Each regression should then only include the individuals who could have been sampled to at least one of the cases the control in question could have been sampled for. The estimated weight for the control in question is the corresponding fitted value from the regression. With this alternative approach, the issue of how to include the caliper matching criterion in the regression vanishes. However, for some values of the matching variables there might be few possible controls and the sampling probabilities will be estimated based on few subjects. As far as I know, this method has not been tested; it is therefore not possible to conclude how sensitive it would be for this kind of problems. It is also possible to extend the KM-weights, for instance by doing a Cox-regression on a sampling indicator with the matching variables as covariates, use the estimated parameters γ from this regression in a Breslow estimator to find $\hat{H}_0(t_i)$, the cumulative baseline hazard and finally estimate p_i with $1-\exp\{-\exp(\hat{\gamma}' y_i)\hat{H}_0(t_i)\}$. The estimated sampling probabilities will now correspond to 1 minus a survival probability where the event is to be sampled as a control. I have not tried this out in practice either, but this extension will have the same problems regarding caliper matching. We could, however, speculate that it might be somewhat better with regard to close matching. In some situation the matching criteria is unnecessary close and estimating weights based on a wider matching criteria may improve the performance of the KM-weights. This can easily be accomplished with our R-package by specifying a wider matching interval.

The NCC-design is based on sampling controls from the cases risk sets, and because of this risk set sampling it is intrinsically linked to Cox-regression and estimation of hazard ratios. However, there exist some possibilities for other regression models and estimation of other interest parameters. Within the framework of the Cox-model, Borgan and Langholz (1993); Langholz and Borgan (1997) showed that NCC-data can be used for estimation of absolute risk and details regarding estimation of cumulative baseline has also been provided (Borgan and Langholz, 1993; Borgan et al., 1995). Ganna et al. (2012) performed estimation of absolute risk for case-cohort and nested case-control data both with unmatched/unstratified and matched/stratified designs. When considering other models, Borgan and Langholz (1997); Zhang and Borgan (1999) generalized Aalen's additive model (Aalen, 1980) to NCC-data and used this to estimate excess and absolute risk. Zheng et al. (2012) derived estimation equations for NCC-data under a class of general additive-multiplicative hazard models (Lin and Ying, 1995). Although, these possibilities exist for NCC-data, the IPW approach is a more unified alternative since weighted analyses parallel to cohort analyses can be performed. Details regarding variance estimation must, however, be sorted out. Chen et al. (2012) utilized the IPW approach and proposed weighted versions of linear transforation models (Cheng et al., 1995) for data from generalized case-cohort designs (Chen, 2001), which include the nested case-control and case-cohort designs, and Samuelsen (1997); Salim et al. (2009) considered parametric models for the NCC-design.

In some situation it can be of interest to estimate the cohort mean of the exposure variables only known for cases and controls. For example, using the vitamin D measurements from the prostate cancer data used in Paper III to estimate the mean vitamin D level for the Norwegian population, possibly within month of blood sampling, could be interesting. Taking a crude mean over cases and controls will in most situations yield a biased estimate due to the over-sampling of cases. Using only controls is probably better, however, a smaller bias in the opposite direction is expected since all cases are excluded. A natural way to do this is therefore to exploit the sampling probabilities by using the Horvitz-Thompson estimator. This will give an unbiased estimate of the population mean, however, this estimator may have a very large variance and as noted earlier it is best to use this estimator when there is a fairly strong negative correlation between the weights and the variable you want to estimate the mean of. This will generally not be the case for NCC-data. With a strong association between the exposure and the outcome, there may be a correlation between the cases weights (which is 1) and the exposure; however the weights and the exposure are usually uncorrelated for the controls. I have conducted some simulation experiments to evaluate the performance of this estimator. As was expected from the general theory, the estimated mean was close to the true mean, however with a large variance. The estimated variance was always much larger than the empirical variance and the reason for this may be too few simulations.

Within the world of causality, it is well known that large weights can cause trouble. I have, however, still not experienced trouble with large weights in connection to the weighted partial likelihood. The reason for this may be that the subjects receiving large weights are the subjects that are followed only for a short period of time, thus having a small probability of being sampled. Those subjects will contribute to the likelihood at few event times, and even though their contribution will be large at the event times they are contributing, it evens out in the overall product over all event times. Another way to see this is from a survey sampling point of view. Then we know that there should be a positive correlation between the sampling probabilities and the variable one wants to estimate the total of, y. With IPW analyses of NCC-data, the individual contributions to the weighted likelihood should play the role of y. They will be small for subjects followed for a short period of time and those subjects are also unlikely to be sampled. They therefore have a small sampling probability, and the positive correlation needed for the H-T estimator is automatically achieved.

Paper III is targeting epidemiologists. They are frequently using the NCC-design and one goal with this thesis was to increase their awareness of this alternative analyzing option for NCC-data. I hope that IPW will appear more tempting when the concept is explained in a more applied setting. Moreover, I hope that the estimation will become simpler with our `R`-package `multipleNCC`, and that this will lead to increasing use of IPW for NCC. The package is documented in Paper IV. It estimates weights and carry out weighted Cox-regressions with a simple one-line call with similar syntax as the `coxph`-function. I hope that this will increase the popularity of IPW for NCC.

References

Aalen OO. A model for non-parametric regression analysis of counting processes. In Klonecki W, Kozek A, and Rosinski J (editors), *Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory*, pages 1–25. Springer Verlag, New York, 1980.

Aalen OO, Borgan Ø, and Gjessing HK. *Survival and event hstory analysis*. Statistics for Biology and Health. Springer, first edition, 2008.

Andersen PK and Gill RD. Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, 10(4):1100–1120, 1982.

Anderson JA. Separate sample logistic discrimination. *Biometrika*, 59(1):19–35, 1972.

Armitage P. The use of cross-ratio in aetiological surveys. In Gani J (editor), *Perspectives in probability and statistics*. Academic Press, Massachusetts, 1975.

Barlow WE. Robust variance estimation for the case-cohort design. *Biometrics*, 50(4):1064–1072, 1994.

Basu D. On the logical foundations of survey sampling. In Ghosh JK (editor), *Statistical information and likelihoods: A collection of critical essays*. Springer-Verlag, New York, 1988.

Bender R, Augustin T, and Blettner M. Generating survival times to simulate Cox proportional hazards model. *Statistics in Medicine*, 24(11):1713–1723, 2005.

Bond GG and Flores GH. Exposure classification in nested case-control studies. *American Journal of Epidemiology*, 120(3):503–503, 1984.

Borgan Ø and Langholz B. Nonparametric estimation of relative mortality from nested case-control studies. *Biometrics*, 49(2):593–602, 1993.

Borgan Ø and Langholz B. Estimation of excess risk from case-control data using Aalen's linear regression model. *Biometrics*, 53(2):690–697, 1997.

Borgan Ø and Olsen EF. The efficiency of simple and counter-matched nested case-control sampling. *Scandinavian Journal of Statistics*, 26(4):493–509, 1999.

Borgan Ø and Samuelsen SO. Nested case-control and case-cohort studies. In Klein JP, Houwlingen HC, Ibrahim JG, and Scheike TH (editors), *Handbook of Survival Analysis*. Chapman and Hall, London, 2013.

Borgan Ø, Goldstein L, and Langholz B. Methods for the analysis of samled cohort data in the Cox proportional hazards model. *Annals of Statistics*, 23(5):1749–1778, 1995.

Borgan Ø, Langholz B, Samuelsen SO, Goldstein L, and Pogoda J. Exposure stratified case-cohort designs. *Lifetime Data Analysis*, 6(1):39–58, 2000.

Breslow NE. Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91(433):14–28, 1996.

Breslow NE and Day NE (editors). *Statistical Methods in Cancer Research: Volume 1 - The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon, first edition, 1980.

Breslow NE, Day NE, Halvorsen KT, Prentice RL, and Sabai C. Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology*, 108(4): 299–307, 1978.

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, and Kulich M. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*, 169(11):1398–1405, 2009a.

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, and Kulich M. Improved Horvitz-Thompson estimation of model parameters for two-phase stratified samples: Applications in epidemiology. *Statistics in Bioscience*, 1(1):32–49, 2009b.

Cai T and Zheng Y. Evaluating prognostic accuray of biomarkers in nested case-control studies. *Biostatistics*, 13(1):89–100, 2012.

Chen HY and Little RJA. Proportional hazard regression with missing covariates. *Journal of the American Statistical Society*, 94(447):896–908, 1999.

Chen K and Lo S-H. Case-cohort and case-control analysis with Cox's model. *Biometrika*, 86 (4):755–764, 1999.

Chen K, Sun L, and Tong X. Analysis of cohort survival data with transformation model. *Statistica Sinica*, 22(2):489–508, 2012.

Chen KN. Generalized case-cohort sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 63(4):791–809, 2001.

Cheng SC, Wei LJ, and Ying Z. Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845, 1995.

Clendenen TV, Lundin E, Zeleniuch-Jacquotte A, Koenig KL, Berrino F, Lukanova A, Lokshin AE, Idahl A, Ohlson N, Hallmans G, Krogh V, Sieri S, Muti P, Marrangoni A, Nolen BM, Liu ML, Shore RE, and Arslan AA. Circulation inflammation markers and risk of Epithelial Ovarian Cancer. *Cancer Epidemiology, Biomarkers and Prevention*, 20(5):799–810, 2011.

Cnattingius S, Hultman CM, Dahl M, and Sparén P. Very preterm birth, birth trauma, and the risk of anorexia nervosa among girls. *Archives of General Psychiatry*, 56(7):634–638, 1999.

Cochran WG and Rubin DB. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 35(4):417–446, 1973.

Cologne J, Preston DL, Imai K, Misumi M, Yoshida K, Hayashi T, and Nakachi K. Conventional case-cohort design and analysis for studies of interaction. *Internation Journal of Epidemiology*, 41(4):1174–1186, 2012.

Cornfield J. A method of estimating comparative rates from clinical data. Application to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, 11(6):1269–1275, 1951.

Cornfield J, Gordon T, and Smith WW. Quantal response curves for experimentally uncontrolled variables. *Bulletin of the International Statistical Institute*, 38(3):97–115, 1960.

Cosslett SR. Maximum likelihood estimator for choice-based samples. *Econometrica*, 49(5): 1289–1316, 1981.

Cox DR. Some procedures connected with the logistic qualitative response curve. In David FN. (editor), *Research papers in Statistics: Festshcrift for J Neyman*. John Wiley, New York, 1966.

Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972.

Crowther MJ and Lambert PC. Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32(15), 2013.

Dahm CC, Keogh RH, Spencer EA, Greenwood DC, Key TJ, Fentiman IS, Shipley MJ, Brunner EJ, Cade JE, Burley VJ, Mishra G, Stephen AM, Kuh D, White IR, Luben R, Lentjes MAH, Khaw KT, and Rodwell SAB. Dietary fiber and colorectal cancer risk: A nested case-control study using food diaries. *Journal of the National Cancer Institute*, 102(9):614–626, 2010.

Day NE and Kerridge DF. A general maximum likelihood discriminant. *Biometrics*, 23(2): 313–323, 1967.

Deville JC and Särndal CE. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992.

Deville JC, Särndal CE, and Sautory O. Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423):1013–1020, 1993.

Floderus B, Persson T, Stenlund C, Wennberg A, Öst Å, and Knave B. Occupational exposure to electromagnetic fields in relation to leukemia and brain tumors: A case-control study in Sweden. *Cancer Causes and Control*, 4(5):465–476, 1993.

Ganna A, Reilly M, Faire U, Pedersen N, Magnussson P, and Ingelsson E. Risk prediction measures for case-cohort and nested case-control designs: An application to cardiovascular disease. *American Journal of Epidemiology*, 175(7):715 – 724, 2012.

Goldstein L and Langholz B. Asymptotic theory for nested case-control sampling in Cox regression models. *Annals of Statistics*, 20(4):1903–1928, 1992.

Grimsrud TK, Berge SR, Haldorsen T, and Andersen A. Exposure to different forms of nickel and risk of lung cancer. *American Journal of Epidemiology*, 156(12):1123–1132, 2002.

Hankinson SE, Willett WC, Colditz GA, Hunter DJ, Michaud DS, Deroo B, Rosner B, Speizer FE, and Pollak M. Circulating concentrations of insulin-like growth factor-I and risk of breast cancer. *The Lancet*, 351(9113):1393–1396, 1998.

Hansen MH and Hurwitz WN. On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14:333–362, 1943.

Hastie TJ and Tibshirani RJ. *Generalized additive models*. Chapman & Hall, London, 2009.

Horvitz DG and Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Hosmer DW and Lemeshow S. *Applied logistic regression*. Wiley Series in probability and mathematical statistics. Wiley, New York, 1989.

Hultman CM, Sparén P, Takei N, Murray RM, and Cnattingius S. Prenatal and perinatal risk factors for schizophrenia and affective psychosis, and reactive psychosis of early onset: Case-control study. *British Medical Journal*, 318(7181):421–426, 1999.

Juul A, Scheike T, Davidsen M, Gyllenborg J, and Jørgensen T. Low serum insulin-like growth factor I is associated with increased risk of ischemic heart disease: A population-based case-control study. *Circulation*, 106(8):939–944, 2002.

Kalbfleisch JD and Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, 7(1-2):149–160, 1988.

Keogh RH and White IR. Using full-cohort data in nested case-control and case-cohort studies by multiple imputation. *Statistics in Medicine*, 2013.

Kim S and De Gruttola V. Strategies for cohort sampling under the Cox proportional hazards model, application to an AIDS clinical trial. *Lifetime Data Analysis*, 5(2):149–172, 1999.

Kulathinal S and Arjas E. Bayesian inference from case-cohort data with multiple end-points. *Scandinavan Journal of Statistics*, 33(1):25–36, 2006.

Kulich M and Lin DY. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, 99(467):832–844, 2004.

Langholz B and Borgan Ø. Counter-matching: A stratified nested case-control sampling method. *Biometrika*, 82(1):69–79, 1995.

Langholz B and Borgan Ø. Estimation of absolute risk from nested case-control data. *Biometrics*, 53(3):767–774, 1997.

Langholz B and Goldstein L. Risk set sampling in epidemiologic cohort studies. *Statistical Science*, 11(1):35–53, 1996.

Langholz B and Jiao J. Computational methods for case-cohort studies. *Computational Statistics and Data Analysis*, 51(8):3737–3748, 2007.

Langholz B and Thomas DC. Efficiency of cohort sampling designs: Some surprising results. *Biometrics*, 47(4):1563–1571, 1991.

Levine RJ, Maynard SE, Qian C, Lim KH, England LJ, Yu KF, Schisterman EF, Thadhani R, Sachs BP, Epstein FH, Sibai BM, Sukhatme VP, and Karumanchi SA. Circulating angiogenic factors and the risk of preeclampsia. *New England Journal of Medicine*, 350(7):672–683, 2004.

Lilienfeld AM and Lilienfeld DE. A century of case-control studies: Progress. *Journal of Chronic Diseases*, 32(1-2):5–13, 1979.

Lin DY and Wei LJ. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078, 1989.

Lin DY and Ying Z. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, 88(424):1341–1349, 1993.

Lin DY and Ying Z. Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *Annals of Statistics*, 23(5):1712–1734, 1995.

Lubin JH. Extensions of analytic methods for nested and population-based incident case-control studies. *Journal of Chronic Diseases*, 39(5):379–388, 1986.

Lumley T. *Complex Surveys: A Guide to Analysis Using R*. Wiley series in survey methodology. Wiley, New York, 2010.

Lumley T, Shaw PA, and Dai JY. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistics Review*, 79(2):200–220, 2011.

Manski CF and Lerman SR. The estimation of choice probabilities from choice based samples. *Econometrica*, 45(8):1977–1988, 1977.

Mantel N and Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

Marinussen T and Scheike TH. *Dynamic regression models for survival data*. Statistics for Biology and Health. Springer, first edition, 2006.

Mark SD and Katki HA. Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (nested) cohort studies with missing case data. *Journal of the American Statistical Association*, 101(474):460–471, 2006.

Marti H and Chavance M. Multiple imputation analysis of case-cohort studies. *Statistics in Medicine*, 30(13):1595–1607, 2011.

Meyer HE, Robsahm TE, Bjørge T, Brunstad M, and Blomhoff R. Vitamin D, season and the risk of prostate cancer: A nested case-control study within Norwegian health studies. *American Journal of Clinical Nutrition*, 97(1):147–154, 2013.

Oakes D. Survival times: Aspects of a partial likelihood. *International Statistical Review*, 49 (3):235–252, 1981.

Øyen N, Markestad T, Skjærven R, Irgens LM, Helweg-Larsen K, Alm B, Norvenius G, and Wennergren G. Combined effects of sleeping position and prenatal risk factors in Sudden Infant Death Syndrome: The Nordic Epidemiological SIDS Study. *Pediatrics*, 100(4):613–621, 1997.

Parsonnet J, Friedman GD, Vansdersteen DP, Chang Y, Vogelman JH, Orentreich N, and Sibley RK. Helicobacter Pylori infection and the risk of Gastric Carcinoma. *New England Journal of Medicine*, 325(16):1127–1131, 1991.

Pischon T, Girman CJ, Hotamisligil GS, Rifai N, Hu FB, and Rimm EB. Plasma adiponectin levels and risk of myocardial infarction in men. *The American Journal of Medicine*, 291(14): 1730–1737, 2004.

Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11, 1986.

Prentice RL and Breslow NE. Retrospective studies and failure time models. *Biometrika*, 65 (1):153–158, 1978.

Prentice RL and Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

Pugh M, Robins J, Lipsitz S, and Harrington D. Inference in the Cox proportional hazards model with missing covariate data. Unpublished report.

R: A Language and Environment for Statistical Computing. R Development Core Team, Vienna, Austria, 2007. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Robins JM, Rotnitzky A, and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

Royston P and Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197, 2002.

Rubin D. *Multiple imputation for nonresponse in surveys*. Wiley, New York, 1987.

Saarela O and Kulathinal S. Conditional likelihood inference in a case-cohort design: An application to haplotype analysis. *The International Journal of Biostatistics*, 3(1), 2007.

Saarela O, Kulathinal S, Arjas E, and Läärä. E. Nested case-control data utilized for multiple outcomes: A likelihood approach and alternatives. *Statistics in Medicine*, 27(28):5991–6008, 2008.

Salim A, Hultman C, Sparén P, and Reilly M. Combining data from 2 nested case-control studies of overlapping cohorts to improve efficiency. *Biostatistics*, 10(1):70–79, 2009.

Salim A, Yang Q, and Reilly M. The value of reusing prior nested case-control data in new studies with different outcome. *Statistics in Medicine*, 31(11-12):1291–1302, 2012.

Samuelsen SO. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84(2):379–394, 1997.

Samuelsen SO, Ånestad H, and Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics*, 34(1):103–119, 2007.

Scheike TH and Juul A. Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics*, 5(2):193–206, 2004.

Scheike TH and Martinussen T. Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scandinavian Journal of Statistics*, 31(2):283–293, 2004.

Self SG and Prentice RL. Asymptotic-distribution theory and efficiency results for case-cohort studies. *Annals of Statistics*, 16(1):64–81, 1988.

Støer NC. Properties of estimators for relative risks from nested case-control studies with multiple outcomes (competing risks), 2010. Master thesis, University of Oslo. `http://urn.nb.no/URN:NBN:no-25646`.

Suissa S, Edwardes MDD, and Boivin JF. External comparisons from nested case-control designs. *Epidemiology*, 9(1):72–78, 1998.

Therneau TM and Li H. Computing the Cox model for case-cohort designs. *Lifetime Data Analysis*, 5(2):99–112, 1999.

Thomas DC. Addendum to: "Methods of cohort analysis: Appraisal by application to asbestos mining" by Liddell FDK, McDonald JC and Thomas DC. *Journal of the Royal Statistical Society: Series A (General)*, 140(4):469–491, 1977.

Tynes T and Haldorsen T. Electromagnetic fields and cancer in children residing near Norwegian high-voltage power lines. *American Journal of Epidemiolgy*, 145(3):219–226, 1997.

Zeng D, Lin DY, Avery CL, North KE, and Bray MS. Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics*, 7(3):486–502, 2006.

Zhang J and Borgan Ø. Aalen's linear model for sampled risk set data: A large sample study. *Lifetime Data Analysis*, 5(4):351–369, 1999.

Zheng M, Lin R, Sun Y, and Yu W. Nested case-control analysis with general additive-multiplicative hazard models. *Applied Mathematics - A Journal of Chinese Universities: Series B*, 27(2):159–168, 2012.

Zheng Y, Cai T, and Pepe MS. Adopting nested case-control quota sampling designs for the evaluation of risk markers. *Lifetime Data Analysis*, 2013. doi: 10.1007/s10985-013-9270-8.

I

II

III