From the DEPARTMENT OF MEDICAL EPIDEMIOLOGY
AND BIOSTATISTICS
Karolinska Institutet, Stockholm, Sweden

EFFICIENT DESIGN AND ANALYSIS OF EXTENDED CASE-CONTROL STUDIES

Bénédicte Delcoigne

Stockholm 2017

All previously published papers were reproduced with permission from the publisher.
Cover drawing by Elsa Struelens.
Published by Karolinska Institutet.
Printed by Eprint AB 2017

ISBN 978-91-7676-643-9

Department of Medical Epidemiology and Biostatistics

Efficient Design and Analysis of Extended Case-Control Studies

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i Föreläsningssal Hillarp, Retzius väg 8, Karolinska Institutet, Solna

Torsdagen den 1 Juni 2017, kl 13.00

av

Bénédicte Delcoigne
MSc

Huvudhandledare:
Professor Marie Reilly
Karolinska Institutet
Institutionen för medicinsk epidemiologi och biostatistik

Bihandledare:
Professor Kamila Czene
Karolinska Institutet
Institutionen för medicinsk epidemiologi och biostatistik

Associate Professor Agus Salim
La Trobe University, Melbourne, Australia
Department of Mathematics and Statistics

Fakultetsopponent:
Professor Sven Ove Samuelsen
University of Oslo, Norway
Department of Mathematics

Betygsnämnd:
Professor Jasper Lagergren
Karolinska Institutet
Institutionen för molekylär medicin och kirurgi

Docent Sara Hägg
Karolinska Institutet
Institutionen för medicinsk epidemiologi och biostatistik

Docent Anna Ekman
Göteborgs Universitet
Institutionen för medicin
Avdelningen för samhällsmedicin och folkhälsa

Stockholm 2017

“In theory there is no difference between theory and practice. In practice there is.”

Yogi Berra, (1925 –2015)

American professional baseball catcher, manager, and coach.

ABSTRACT

The nested case-control design is widely used in epidemiology for its efficiency, as it combines the advantages of both cohort and case-control designs. This design is an extension of the matched case-control design, where the matching variable is the time of occurrence of the outcome. Consequently, the nested case-control data are usually analysed with conditional logistic regression; however, this analysis suffers from various limitations.

Several authors have developed novel statistical methods for alternative analyses of nested case-control data using basic information from the underlying cohort. Among these methods, one approach consists of ignoring the matching, weighting the sampled individuals to recover a representation of the underlying cohort and analysing the data by maximising a weighted partial likelihood. This method can be considered when two conditions are fulfilled: 1) the sampling was performed in a well-defined underlying cohort for which basic information is available, and 2) the exact sampling procedure is known.

This thesis aimed to refine and extend the scope of the weighted likelihood approach in nested case-control data analysis by investigating the advantages of this method as an alternative to the traditional conditional logistic regression in several situations. The reuse of nested case-control data to address a research question regarding a new outcome, the calculation of absolute risk, the mitigation of the problem of overmatching, the maximisation of the data exploitation in case of clustered data and the analysis of subgroups of nested case-control data were addressed in this thesis. While Studies I and III were motivated by an actual epidemiological question for which data were available, simulation studies were the main approach used in Studies II and IV.

Reusing nested case-control data to address a research question regarding another outcome was the central point of interest in Study I. Addressing an epidemiological question regarding the risk factors for contralateral breast cancer, for which data on contralateral breast cancer case patients were available, the feasibility of reusing nested case-control data from a previous study as the control dataset was studied. Practical aspects of the approach were highlighted, such as the consequences of reusing data which have narrow inclusion criteria, the restriction in the choice of the type of weights which can be calculated and the importance of having information on censoring dates for controls. In addition, we found that an imperfect reconstruction of the study base led to similar estimates in the analysis compared to an appropriate study base reconstruction; moreover, we confirmed that using unstratified weights (in cases of stratified sampling) provided similar exposure estimates than stratified weights, provided that adjustments were made on the confounder variables which drove the sampling. We also confirmed that using a naïve unweighted method instead of an appropriate method led to biased estimates.

Absolute risk estimation was studied in Study II. Two methods were compared with both simulation studies and a real data application. The ability of each method to provide valid absolute risk estimates was investigated, in particular in cases of matched study designs.

Both the Langholz-Borgan and weighted methods provided valid estimates in most situations, the latter showing slightly higher levels of precision than the former. In case of fine matching, the Langholz-Borgan method was more prone to be biased than the weighted method and had larger standard errors.

In Study III, we handled nested case-control data, which had been collected to address an epidemiological question regarding how radiation therapy and smoking interact in their association with lung cancer in female breast cancer patients. Data on paired organs (breast and lungs) were collected for exposure and outcome variables, which provided clustered data at the individual level. The collected data was also characterised by the problem of overmatching which arose at the design stage. Using weighted partial likelihood allowed mitigation of the problem of overmatching and better exploited the collected data, compared to conditional logistic regression. In addition, a further advantage of the weighted approach was to enable calculating the absolute risk for a lung to develop cancer given the radiation therapy dose received for breast cancer treatment and the smoking habits of the patient.

In Study IV, we compared the conditional logistic regression and weighted likelihood methods in terms of validity and efficiency of nested case-control data subgroup analyses, with subgroups defined by different variables measured at baseline. All investigated subgroup analyses provided valid estimates with both analyses. The advantages of weighted likelihood compared to conditional logistic regression were highlighted for the estimate's precision. In addition, we showed that the weighting system enabled, on average, the reconstruction of the correct number of individuals at risk over time, for the whole cohort and in subgroups.

In conclusion, the weighted likelihood approach showed several advantages compared to the traditional conditional logistic regression in nested case-control data analysis, which reinforces, refines and extends what has been previously shown in the literature.

LIST OF SCIENTIFIC PAPERS

I. **Delcoigne B**, Hagenbuch N, Schelin ME, Salim A, Lindström LS, Bergh J, Czene K, Reilly M. Feasibility of reusing time-matched controls in an overlapping cohort. *Stat Methods Med Res.* 2016 Sep.
pii: 0962280216669744. [Epub ahead of print].

II. Salim A, **Delcoigne B**, Villaflores K, Koh WP, Yuan JM, van Dam RM, Reilly M. Comparisons of risk prediction methods using nested case-control data. *Stat Med.* 2017;36(3):455-465.

III. **Delcoigne B**, Colzani E, Prochazka M, Gagliardi G, Hall P, Abrahamowicz M, Czene K, Reilly M. Breaking the matching in nested case-control data offered several advantages for risk estimation. *J Clin Epidemiol.* 2017;82:79-86.

IV. **Delcoigne B**, Støer N, Reilly M. Valid and efficient subgroup analyses using nested case-control data. Submitted.

CONTENTS

LIST OF ABBREVIATIONS

BRCA1/2	Breast Cancer gene 1/2
CBC	Contralateral Breast Cancer
gam	General additive model
glm	General linear model
HDL	High-Density Lipoprotein
IPW	Inverse Probability Weighting
NCC	Nested case-control
L-B	Langholz-Borgan
SAMBAL	Molecular epidemiology of secondary lung cancer study
SBP	Systolic Blood Pressure
TNM	Cancer staging: Tumour-lymph Node-Metastasis

1 INTRODUCTION

Choosing the best observational study design in epidemiology is driven by both the research question and ethical and practical considerations, including the resources and time available for data collection. Two classical study designs are the cohort studies, the paradigmatic gold standard, and case-control studies, which have advantages in cost-efficiency, and are preferred when the outcome under investigation is rare. To improve these study designs, alternative sampling methods have been suggested. This thesis examines one of them: the nested case-control design, which combines the benefit from the time aspect of cohort studies and the significant savings in cost and time of case-control studies.

Conditional logistic regression is the traditional statistical approach used to analyse nested case-control data. This approach, however, has several limitations. For example, the data cannot be reused to address a new research question regarding another outcome, or to estimate the absolute risk of developing the outcome of interest. Since the late 1990s, alternative statistical methods have been developed to analyse nested case-control data. These non-traditional methods ignore the matching between cases and controls, and use limited information from the underlying cohort.

Among these methods, the Inverse Probability Weighting (IPW) method consists of weighting the individuals and analysing the data by maximising a weighted partial likelihood. The weights are the inverse of the probability for each individual to be sampled into the nested case-control study and aim to reconstruct the number of individuals at risk in the full cohort. Simulation studies have validated the IPW method and have consistently shown that it provides more precise exposure estimates, compared to the conditional logistic regression method. Although the method has long been available, it is still largely ignored by medical researchers and is uncommonly used by biostatisticians and epidemiologists.

The main goal of this thesis is to refine and extend the scope of this weighted partial likelihood method and make it more accessible to the research community by providing guidelines illustrated with applications to clinical datasets. The contribution of this study may encourage a shifting of the standard of statistical analysis of nested case-control studies from conditional logistic regression to the more powerful and versatile weighted partial likelihood method.

2 BACKGROUND

2.1 Cohort and case-control studies

In epidemiology, for ethical and practical reasons, observational studies are far more often conducted than randomised control trials. The most popular observational designs have long been and still are cohort and case-control study designs, with the former considered as the gold standard when randomisation is not possible.[1,2]

A cohort is a well-defined group of study individuals who are followed from inclusion in the study until the time of occurrence of an event of interest (i.e. outcome) or censoring. Some cohort members will be exposed to the risk (or protective) factor under investigation, while other will remain unexposed. How exposure is related to time until outcome is analysed by contrasting the unexposed subjects with the exposed subjects. In contrast, the classical case-control design consists of sampling a subset from a larger population, which can be a cohort. Individuals experiencing the outcome, i.e. cases, and controls, who do not experience the event, are sampled separately. Usually all cases are selected, and controls randomly sampled among the non-cases. The sampling of controls can be a simple random sampling, but more often involves stratification (i.e. matching). How the outcome is related to the exposure is analysed by contrasting the case and control individuals.

Beyond the difference in the way the participants of the study are selected, the cohort design enables the modelling of the time aspect of the occurrence of an outcome, while the case-control study design is far more cost-effective but usually does not take the time into account.[1,2]

2.2 Time-to-event analysis of cohort studies

2.2.1 Basic concepts

To set up a cohort of individuals who are followed over time requires a precise definition of the outcome/event of interest, a precise definition of the start and the end of follow-up time period, a relevant choice for the time scale and precise inclusion/exclusion criteria for the individuals who enter the cohort. Subjects who do not experience the event during follow-up are censored. The censoring time is usually either end of follow-up, or the time when the individual was lost to follow-up for other reasons (such as death, emigration or withdrawal from the study). The observed follow-up time is the minimum between censoring time and event time. Moreover, the individuals are not always followed from time zero, but can sometimes be observed from some later time. This often happens, for example, when age is used as time scale. When individuals enter the cohort after time zero is referred to as delayed entry.

The time-to-event or survival time T is the central variable in cohort analysis and the probability density function $f(t)$ describes how the variable T is distributed. This function is used to further express the distribution function of T:

$$F(t) = \mathrm{P}(T < t) = \int_0^t f(u)du \qquad (2.1)$$

which in turn is related to one of the central functions of interest in survival analysis: the survivor function

$$S(t) = 1 - F(t) \qquad (2.2)$$

which expresses the probability to have survived until time *t*.

Another central function in survival analysis is the hazard function *h(t)* which expresses the risk of experiencing the event of interest at time *t*, provided one has survived until that time. The two central functions *h(t)* and *S(t)* are related to each other by means of the cumulative hazard function

$$H(t) = \int_0^t h(u)du \qquad (2.3)$$

The relationship between these central concepts are derived with simple algebra [3]:

$$H(t) = –\log S(t) \qquad (2.4)$$

The cohort design is illustrated in Figure 2.1. The members of this small cohort enter the study from the beginning (i.e. no delayed entry) and are followed over time.

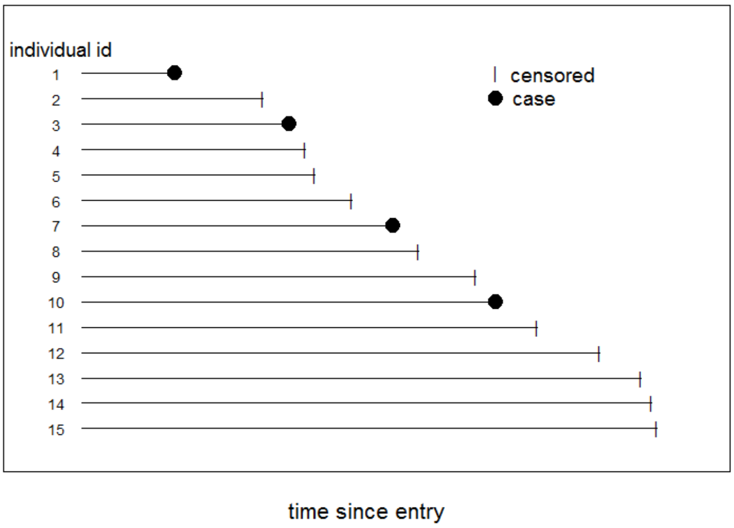

Figure 2.1: Cohort of 15 individuals with ordered event and censoring times.

2.2.2 Modelling survival data with Cox proportional hazards model

The most usual way to describe the hazard function for an individual *i* at time *t*, is to express this hazard as a product of two functions: a baseline hazard ($h_0(t)$), which is multiplied by another function which includes the variables which characterise individual *i*. In the Cox proportional hazards model, this expression is

$$h_i\,(t|X_i, Z_i, \beta, \gamma) = h_0(t)\ exp(\beta'X_i + \gamma'Z_i) \qquad (2.5)$$

where X_i and Z_i are the vectors of covariate(s) and confounder(s) for individual i, and β and γ are the vectors of the corresponding coefficients. The baseline hazard function $h_0(t)$ can vary over time but this function does not need to be specified in the Cox proportional hazards model, which is the reason why this model is referred to as a semi-parametric model. As the ratio of the hazard function of two individuals does not include the baseline hazard (which is cancelled out in the ratio), the hazard ratio is constant over time: a feature which gives its name to the model, i.e. the proportional hazards model.

Ordering the event times t_i, the vectors of coefficients β and γ are estimated by maximising the following partial likelihood [4]:

$$L(\beta, \gamma) = \prod_{t_i} \frac{\exp[\beta' X_i + \gamma' Z_k)]}{\sum_{k \epsilon R_i} \exp[\beta' X_k + \gamma' Z_k)]} \tag{2.6}$$

where R_i is the group of individuals who are still at risk just before time t_i and is called the risk set at time t_i.

2.2.3 Comments on the risk set concept

The central concept in the Cox regression analysis is the concept of risk set. The data which is used in each contribution of the product (2.6) is represented in Figure 2.2. The figure highlights the concept of risk set as well as related features.

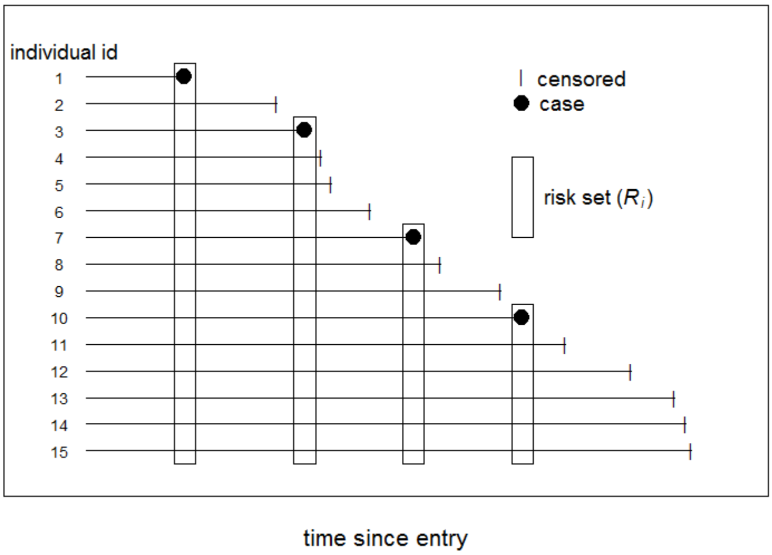

Figure 2.2: Cohort of 15 individuals with ordered event and censoring times and the risk set R_i at each event time.

- In a Cox regression the only interest lies in the number of individuals who are present in the risk sets at the exact event times, disregarding what could have happened just before or after;
- The ranking of the time is only used to identify who is in the risk sets, but even this ranking is not formally used in Equation 2.6;

- For case *i*, the number of terms in the denominator of Equation 2.6 is as large as the number of individuals in the risk set R_i, decreasing thus with time.

This last feature may give some clue to a possible sampling strategy within the risk sets, which could reduce the volume of data to be collected and used for analysis.

2.3 Case-control studies and logistic regression analysis

2.3.1 Basic concepts

When the outcome of interest is rare, the cohort is not the most efficient way to address an etiologic question as it will require collecting data on a high number of individuals over a long period of time in order to obtain a sufficient number of cases. When an exposure is expensive to collect (as with biomarkers), collecting data for all of the individuals in a cohort is very costly. In these two situations, the case-control design would be much more efficient.

In addition to defining the outcome/event of interest with precision, the sampling procedure of the control individuals requires careful consideration in order to enable valid inference. Cases and controls have to be representative of one single population (i.e. the study base) but are not required to be sampled within a cohort. Indeed, most of them are not, but are rather sampled within a dynamic underlying population in which the members may vary over time.[5,6]

In contrast to the cohort design, time (to event) is not the main area of focus in the case-control design. Time is however considered in the sampling: the source population is followed over a particular period of time and, in addition, the sampling of the controls can be chosen to be performed at different time points. It can be performed at the beginning of the study period, concurrently (by matching on time with the cases), at the end of the study period, or during the entire study period.[6-8] When the sampling is done within a cohort (the situation of interest in this thesis), the three first choices are respectively referred to as inclusive sampling, incidence density sampling and exclusive sampling.[6,7]

2.3.2 Modelling and analysing case-control data

Case-control data are usually described using a logistic model, and analysed using logistic regression. Confounding is adjusted for by including the confounders as covariates in the regression model. Such analyses provide odds ratios which measure the association between the case/control status and the exposure. The odds ratio estimates either a risk ratio, a rate ratio or an odds ratio, depending on the sampling design used for selecting controls and the source population (cohort or dynamic).[7]

Case-control studies have long been, and still are, criticized as being less reliable than studies utilising a cohort approach, despite numerous developments and scientific papers demonstrating the strong links between the two designs as well as similarities of the parameters which are estimated from both designs.[6-7,9-11]

2.3.3 Individually matched case-control studies

Matching on factors which are possible confounders is commonly used in case-control studies. The purpose of matching is to improve efficiency by achieving a better balance in the number of cases and controls in the confounder strata. On the other hand, as the cases and controls are (artificially) made more similar for both the matching factor and the exposure, the process of matching introduces a bias which has to be controlled for in the analysis.[12]

When individual matching is performed on *Z*, matched sets are obtained in which one case is associated with m controls and the individuals in a set share the same value of the matching variable *Z*. Such data is analysed by conditional logistic regression characterised by the following likelihood [13]:

$$L(\beta) = \prod_i \frac{\exp[\beta' X_i]}{\sum_{k \epsilon S_i} \exp[\beta' X_k]} \tag{2.7}$$

where S_i is the set including case *i* and the *m* matched controls. The form of this likelihood is reminiscent of the partial likelihood (2.6) used for analysing cohort data, but γ is not estimated in the likelihood as the confounder *Z* is used as matching factor. The link between the analysis of cohort data and the analysis of matched case-control data can be highlighted with the likelihoods 2.6 and 2.7. Each contribution in both likelihoods includes one single ‘set’ (an entire risk set R_i for the cohort; a matched risk set S_i for the matched case-control). Time is not formally included in any of the analyses.

2.4 Nested case-control design and its traditional analysis

Optimising the efficiency of epidemiological research is a continuous challenge, leading to the development of new designs and new statistical methods to analyze the collected data. The nested case-control design was proposed by Thomas in 1977 [14] and aims to combine the benefit from the time aspect of the cohort design and the significant savings in cost and time of the case-control design.[2,15-17]

2.4.1 Basic concept

Within a well-defined cohort, the nested case-control design includes all cases but only a random sample of controls: at each failure time among all individuals who are at risk at that time, a defined number m of controls is randomly sampled. This incidence density sampling is also called ‘risk-set sampling’ as indeed, the sampling is performed within each risk set as illustrated in Figure 2.3.[2,6,17]

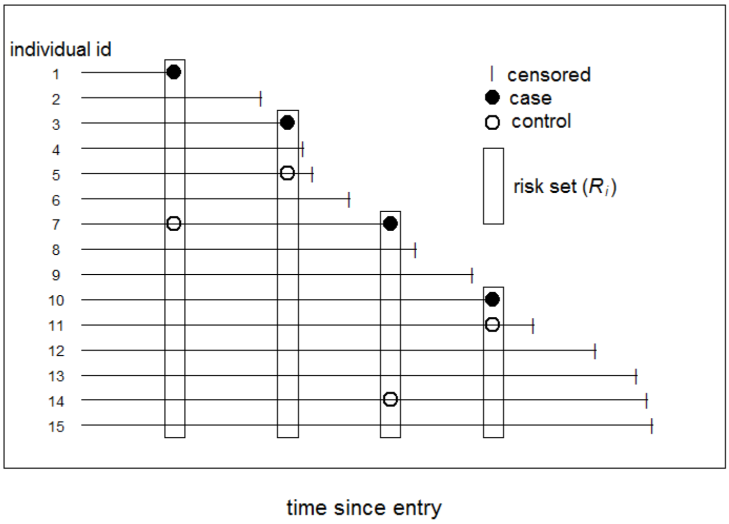

Figure 2.3: Risk set sampling of nested case-control data with a single control per case (*m*=1) in a cohort of 15 individuals.

2.4.2 Different ways to consider the nested case-control design

The nested case-control design is an extension of the case-control design, in particular the individually matched case-control design, where the matching variable is the time of occurrence of the outcome. Hence, statistical methods developed to analyse matched case-control data can be used.

As the sampling is performed within the risk sets, it emphasises the link between the nested case-control and the cohort designs. Instead of including the whole risk set at each event time, only a few individuals are included. As for the cohort design, the same individuals can participate in several sampled risk sets (since individuals are randomly sampled, they can be sampled several times), and controls included in a sampled risk set can later become cases during their follow-up. These two features are typical of cohort design.

As the design associates aspects of both cohort and case-control designs, it is interesting to see that some authors [1,6-8] present this as an extension of case-control design and others [15,17,18] as a cohort sampling method. These two points of view emphasise once more the links between cohort and case-control designs, which are still too often presented as two radically different approaches when performing epidemiological research. This observation was made by Miettinen in 1982 [9] and is still valid today.

2.4.3 Nested case-control design in epidemiological research

Nested case-control design is regularly used in epidemiology but a simple literature search on PubMed shows that this design is still far less used than cohort and case-control designs. There are, however, numerous published cohort studies which might have been more cost and time efficient if they had sampled a nested case-control study within their cohort.

On the other hand, it seems that the concept of nested case-control sampling is poorly understood in the research community: in a PubMed search with the only search criterion ‘nested case-control[Title]’, at least three hits among the first 10 were misusing the term ‘nested case-control’.[19-21] When a study claims to use a nested case-control design, it regularly means that the case-control study was sampled within a cohort without using the risk set sampling strategy.[19-22] Unfortunately, such confusion does not help to promote the correct use of the design.

2.4.4 Traditional statistical analysis: the conditional logistic regression

Nested case-control data, considered as a sample within the cohort, can be described in the proportional hazards framework. Using the proportional hazards model (Equation 2.5) to express the association between the time-to-event and the covariates, the classical approach for estimating β and γ with nested case-control data is to maximise a partial likelihood similar to (2.6) where the sampled risk sets R'_i are used instead of the complete risk sets R_i, or, similar to (2.7) where R'_i are the matched risk sets [14]:

$$L(\beta,\gamma) = \prod_{t_i} \frac{\exp[\beta' X_i + \gamma' Z_i]}{\sum_{k \epsilon R'_i} \exp[\beta' X_k + \gamma' Z_k]} \tag{2.8}$$

where R'_i is the sampled risk set for case i. If, in addition to time, the nested case-control sampling included the confounder(s) as matching factor(s), the likelihood (2.8) will not include the vector γ as in the likelihood (2.7). In practice, this is handled in statistical software by either using a conditional logistic regression or a Cox regression stratified on the sets which include the case and its sampled controls.

To ensure valid inference, the nested case-control design requires that the incomplete data arising from the subsampling of the cohort must be missing at random. This means not only that controls have been randomly selected within the strata defined by the matching variables but also that, conditional on the complete covariates, any missing exposure information does not depend on the unobserved value.[17,18,23] If the condition of ‘missing at random’ is fulfilled, the analysis will provide cost-efficient unbiased estimates of hazard ratios for measured risk factors under the proportional hazards model.[2]

As the nested case-control data analysis is performed with fewer data, a loss in power is expected. Defining the relative efficiency as the ratio of the variance of the parameter estimated from the full cohort to the variance estimated from the nested case-control design, it has been shown that, in the case of one covariate, this ratio follows the $m/(m+1)$ rule under the null hypothesis, with m the number of controls in a matched set. If only one control is sampled for each case, the relative efficiency is ½, implying that the variance obtained with the nested case-control design is twice as large as the one obtained from the cohort.[17] When the regression coefficients depart from zero (i.e. under the alternative hypothesis), or when multiple covariates are included in the model, this rule does not strictly apply any more and efficiency can be either reduced [17] or increased.[24]

2.4.5 Limitations of the nested case-control design and its analysis

While conditional logistic regression provides unbiased results and is easily conducted for a nested case-control study, the matching on time of cases and controls implies that:

- the analysis is valid for this particular time scale;
- the controls cannot be readily reused to address another outcome;
- the design is not optimal for estimating absolute quantities such as baseline hazard and absolute risk;
- analysis of data collected on paired organs (e.g., eyes, lungs, breasts) has to be restricted to reduced data.

These limitations have long been regarded as weaknesses of the nested case-control design,[2,16,17,25] but have also stimulated the development of other statistical approaches. In addition to these limitations, which are mainly related to the use of conditional logistic regression, other issues arise from the sampling strategy and would be avoided with a cohort design. These issues include the problem of overmatching at the design stage and the use of post-hoc subgroup analyses.

As the work presented in this thesis was motivated by the need to overcome these limitations and issues, they are described in more detail in the next section.

2.4.6 Limitations and issues in nested case-control studies

2.4.6.1 Re-using nested case-control data

2.4.6.1.1 Motivation

Large and well-defined cohorts are regularly used to address research questions: national/regional health and population registers (Swedish Registers [26]) are used to select cases for a specified outcome and subsequently sample relevant controls using sampling methods (risk sets or others), before collecting more expensive data such as those retrieved by review of medical charts or administering questionnaires. There is an interest in reusing these valuable data to answer other research questions. In other types of large cohorts including biobank initiatives (UK Biobank,[27] Life Gene [28]), large amount of data, often expensive to gather (especially molecular and genetic information), are stored. There is now an increasing interest in utilising these data efficiently as well as for several purposes.

When data are not reused, researchers miss the opportunity to make more efficient use of their data resources. For example, a study in which a series of nested case-control studies were designed to address several research questions in the same underlying cohort,[29,30] would have been more cost- and time- efficient if collecting less control data and reusing the data to address the different research questions.

On the other hand, when previously collected data are reused to address new research questions, they are often analysed with naïve statistical methods.[31,32] Lin et al. [33] showed that case-control association studies often misuse statistical methods when utilising data to analyse secondary phenotypes. In nested case-control studies, as in any regular matched case-control study, the control data is a biased subset of the population and cannot be readily reused to address a new research question. Special attention is needed in order to correct for the sampling bias. Yung and Lin [34] showed in case-control studies that naïve methods, including the following, can lead to severe bias: analysis of the combined sample of cases and controls ignoring the case-control ascertainment, or adjusted for the case-control status as an additional variable, as well as the case-only or control-only analyses.

2.4.6.1.2 Weighting the observations to reuse case-control data

As the cases are overrepresented compared to the non-cases, any case-control sample is biased by definition. Depending on the strategy used to sample the controls, the control data in an unmatched design, may or may not be a biased sample of the population, while the sample of controls in a matched design will be biased. As the nested case-control sample is matched design on time, the same consequence applies.

Reusing (matched) case-control data is feasible provided the sampling bias is corrected for. Reilly et al.[35] addressed such reuse of case-control data using a weighted approach. From the survey literature, it is known that using the sampling probabilities which lead to the available data is the key to compensating for biased sampling schemes.[36,37] Reilly et al [35] showed how control or case-control data can be reused to address a new research question by weighting the individuals by the inverse of their sampling probabilities and, therefore, making them representative of the study population.[35] This inverse probability weighting (IPW) approach can easily be conducted with a straightforward weighted analysis implemented with any standard statistical software.

2.4.6.1.3 Line of thought

In the approach of Reilly et al.,[35] as the case and control observations were essentially reweighted to construct an unbiased cross-sectional representation of the population, the results apply to cross-sectional data only, and not to nested case-control studies. However, the simplicity of the IPW approach is appealing and inspired the same [38] and other authors [39] to develop such an approach for nested case-control data.

2.4.6.2 Absolute risk estimation

2.4.6.2.1 Motivation

Risk prediction is currently used by practitioners and public health authorities to make decisions on medical treatments, patient follow-up and healthcare regulations. Risk prediction models are usually constructed from data collected in cohort studies such as the famous Framingham Heart study of cardiovascular risk factors.[40-43]

2.4.6.2.2 The Breslow estimator in cohort studies

When using a cohort design and analysing the data in the framework of the Cox model, the Breslow estimator is the key estimator which allows calculating the cumulative baseline hazard H_0, and hence calculating absolute risk [3,44]:

$$H_0(t) = \sum_{i} \frac{I\,(t_i \leq t)}{\sum_{k \in R\,i} \exp[\beta' X_k + \gamma' Z_k]} \tag{3.1}$$

Using $H_0(t)$ obtained above together with an estimate of the vectors β and γ and the equations 2.2 and 2.4, the absolute risk (F_i) for an individual i to develop the outcome at time t is given by

$$F_i(t) = 1 - \exp\,(-H_0(t)\ \exp\big[\beta' X_i\ + \gamma' Z_i\ \big]) \tag{3.2}$$

where X_i and Z_i are the vector of the individual's covariates.

2.4.6.2.3 Adapted Breslow estimator for nested case-control data

As cases are over-represented in nested case-control data, the Breslow estimator can not be readily used to estimate the cumulative baseline hazard $H_0(t)$.

Langholz and Borgan [45] proposed a method to estimate absolute risk from nested case-control data and developed an estimator for the cumulative hazard in the context of the proportional hazards model. Their estimator of the cumulative baseline hazard is similar to the Breslow estimator with an additional time-dependent weight $w(t_i)$ in the denominator:

$$H_0(\mathrm{t}) = \sum_{i} \frac{I\,(t_i \leq t)}{\sum_{k \in R' i} \exp[\beta' X_i\ + \gamma' Z_k]\, w(t_i)} \tag{3.3}$$

with R'_i the sampled risk set at time t_i, $w(t_i)$ a time dependent weight which is the inverse of the sampling fraction in the cohort's risk set R_i: $w(t_i) = R_i/(m+1)$, and β and γ the estimated coefficients in the traditional conditional logistic regression analysis.

Despite the availability of the Langholz-Borgan [45] method developed almost two decades ago, few epidemiological nested case-control studies have used their estimator.[46] Studies which have used population-based nested case-control data to construct their absolute risk model are nevertheless not uncommon [47-49]; however, the nested case-control data in these studies provided the estimates for the risk factors (i.e. the vectors β and γ) while absolute measures were derived by combining the risk estimates with incidence rates from population register data.

The Langholz-Borgan [45] approach uses basic information from the cohort to supplement the nested case-control data. The denominator contribution of the estimator is upweighted, correcting for the over-representation of cases in the nested case-control data set. However, in case of additional matching (on another variable than time), it remains unclear how the Langholz-Borgan method can be accommodated, as it uses conditional logistic regression to estimate the regression coefficients. In such a matched situation, it will not be possible to estimate the coefficients γ of the matching factors. The authors argued that the method could easily be extended to nested case-control sampling involving matching factors, but they did not explain it in detail. Ganna et al. [50] used this method and showed that it did not give reasonable estimates when the nested case-control sampling involved matching factors other than time.

2.4.6.2.4 Line of thought

If part of the limitation in estimating absolute risk with nested case-control data is overcome with the Langholz-Borgan estimator, the use of conditional logistic regression remains an obstacle to achieving reliable estimates for the matching variables and hence an absolute risk estimate for matched design. How this estimator can be accommodated in case of matching is worth addressing.

2.4.6.3 Clustered data

2.4.6.3.1 Motivation

When the research question concerns paired organs (e.g., lungs, breasts, eyes) for each of which exposure and outcome measurements are available, the statistical analysis has to take care of the clustering. This can be addressed with a cohort design by using a frailty model approach, which is an extension of the Cox model.[51] In this approach, the survival model includes both fixed and random effect terms, where the fixed effect term comprises the observed portion of the model and the random effect term, or frailty term, accounts for the unexplained heterogeneity in the model: in other words, it accounts for the correlation within the clusters.

For nested case-control designs, however, the clustered data cannot readily be handled by conditional logistic regression, as there are two levels of clustering: the set which comprises (at least) two individuals (case and control(s)) and then the individual who has a pair of organs. Facing such situation, researchers usually solve the problem by using the cases and their ipsilateral control [52] or adopt a case-only design [53] in which the healthy organ in the pair is used as the control for the unhealthy one.

2.4.6.3.2 Breaking the matching in (usual) case-control studies

As the issue of this double level of clustering arises due to the matching of cases and controls, tackling the question of how/if it is possible to perform valid unconditional analysis with matched data is worth considering.

Performing valid analysis of matched data requires the matching variables to be considered in the analysis, but does not strictly require matched analysis (such as conditional logistic regression).[12] This implies that ignoring (or 'breaking') the matching can be considered, i.e.

ignoring the formal link between the case and the matched controls in the analysis. When the matching is ignored, valid analysis must include the factors which were used for the matched sampling as additional variables in the unconditional logistic regression model.[12,54] Using such an approach can lead to more efficient analysis than the matched analysis, for example, when sets are suffering from missing data [12,55]; however, the method works well if the matching factor included in the model does not have too many strata (i.e. fine matching), otherwise the unconditional logistic regression would lead to biased results.[56]

2.4.6.3.3 Line of thought

If this could be an approach for the usual case-control studies, the situation, however, seems more complicated for nested case-control data: the number of strata is the same as the number of cases and will lead to the problem of numerous strata mentioned above. However, the fact that breaking the matching will remove the clustering at the set level and would allow the information gathered for the two organs to be used will retain interest in this approach.

2.4.6.4 Overmatching

2.4.6.4.1 Matching versus overmatching

When the distributions of the potential confounders are substantially different in cases and controls, the matching aims to balance cases and controls within the strata defined by confounders and, in this way, to improve the efficiency of the stratified analysis.[1,12,54,57] However, if the cases and the controls are matched on a variable which is correlated to the risk factor under study, and this risk factor is not an independent risk factor or part of the causal mechanism, the problem known as ‘overmatching’ can be encountered. In such a situation, the efficiency of the analysis can be much reduced due to cases and controls being too similar for their exposure and many sets not contributing in the stratified analysis.[1,54,58-60]

2.4.6.4.2 Mitigating overmatching

It has been argued that for case-control studies suffering from overmatching, pooling the data in several strata will improve the efficiency.[58] The gain from this approach will be more important if the odds ratio is expected to be large and also depends on the exposure prevalence among the controls.[58]

2.4.6.4.3 Line of thought

To my knowledge, this topic has never been addressed in studies using the nested case-control design. However, it can be hypothesised that breaking the matching in nested case-control data could provide some advantages when overmatching is present as it would also benefit from making use of all individuals included in the study.

2.4.6.5 Subgroup analyses

2.4.6.5.1 Motivation

While a research study is designed to answer a specific research question, it is common for investigators to conduct subsequent analyses of subgroups in order to investigate associations in more detail. Some recent examples for the nested case-control design are the studies conducted by Devore et al.,[61] Kim et al., [62] Boursi et al. [63] and Liu et al.[64]

There is an abundant literature on subgroup analyses for randomised clinical trials.[65-70] The general advice, however, is to restrict subgroup analyses as much as possible, as the main issue is their overuse and over interpretation, in addition to the underuse of appropriate statistical tests for interaction.

There is little doubt that the situation is similar with observational studies. While this study is in line with the advice given above and does not wish to promote an excessive use of subgroup analyses, the main objective was to investigate whether such subgroup analyses are valid per se when they are performed with nested case-control data.

The subgroups can be defined by the outcome (for example, when studying breast cancer, a subgroup could be defined by the histology of the breast cancer (e.g., ductal breast cancer, lobular breast cancer), or the TNM stage (e.g., breast cancer with TNM stage <4)), and by a covariate which could be an independent risk factor, a confounder or an effect modifier (for example, in studying menopaused women, or BRCA1/2 carriers, or smokers, etc.).

In cohort studies, defining a subgroup by the outcome is equivalent to addressing a new research question in the same cohort. Defining a subgroup by a covariate is not a problem either, as it involves a redefinition of the inclusion/exclusion criteria. In nested case-control studies, since the sampling of the participants is related to the outcome of interest (and perhaps also to matching variables), the data at hand is not a representative sample of the population of interest. A subgroup selected from this sample will also be a non-representative sample of the sub-population of interest and, therefore, could generate invalid estimates.

Likewise, defining a nested case-control subgroup by the outcome is also equivalent to addressing a new research question. This should not raise any problems: any set where the case does not correspond to the new definition will be removed, reducing the data at hand to answer the new question (unless the removed subjects are reused, which can be an option). Defining a nested case-control subgroup by a covariate which was used for the sampling (i.e. which was a matching factor) should not raise any problems either, as the sampling was stratified on the variable defining the subgroup. But if the covariate used to define the subgroup was not a matching factor, it is unclear to what extent such sub-group analyses could be valid.

2.4.6.5.2 Line of thought

To the best of my knowledge the validity of subgroup analyses of nested case-control data has not been addressed. Moreover, the main concern facing such subgroup analysis was that it is not known who is being analysed and who these analysed individuals represent.

A second concern is about the efficiency of the analyses. Should subgroup analysis be valid with nested case-control data, it is clear that conditional logistic regression will be inefficient: the definition of a subgroup will restrict the analysis to sets that, by chance, have a case and at least one control in the defined subgroup. The number of sets which will be excluded from the analysis will depend on the prevalence of the subgroup but also on how this variable correlates with all others, including the outcome.

Once again, an analysis that ignores the matching and allows using all individuals belonging to the defined subgroup should have advantages over the conditional logistic regression.

2.4.6.6 In conclusion

Two main ideas guided the search for potential solutions for the mentioned limitations: breaking/ignoring the matching, which enables using more data because cases and controls are no longer tied in sets, and weighting the individuals, which permits the recovery of the study population.

In this thesis, a weighted partial likelihood method was used, which resulted from statistical developments inspired by these ideas. The method, also called Inverse Probability Weighting (IPW) method, started to be developed in the 1990s and is presented in the Methods section (Chapter 5).

3 AIMS

The overall aim of the thesis was to refine and extend the scope of the weighted partial likelihood method in nested case-control data analysis by investigating the advantages of the method as an alternative to the traditional conditional logistic regression. We addressed the following themes: reusing data (Studies I, III and IV), estimating absolute risk (Studies II and III), solving a problem of overmatching at the design stage (Study III), analysing clustered data (Study III) and analysing subgroups (Study IV).

The specific methodological and statistical aims included:

1. To appropriately reuse nested case-control data to investigate a new outcome within the same underlying cohort (Study I).

2. Compare two methods of estimating absolute risk with nested case-control data, and investigate the ability of each method to provide valid estimates for matched study designs (Study II).

3. Investigate the advantages of weighted partial likelihood methods compared to conditional logistic regression for analysing overmatched and clustered nested case-control data (Study III).

4. Compare conditional logistic regression and weighted partial likelihood methods in terms of validity and efficiency of subgroup analyses of nested case-control data (Study IV).

Two of these studies (Studies I and III) were motivated by epidemiological questions for which the collected data gave the opportunity to address additional methodological/statistical questions. The choice was made to focus on these latter questions.

4 MATERIALS AND CONTEXT OF THE STUDIES

The methodological work in Studies I and III was motivated by epidemiological research questions for which data were available, and for which both the epidemiological questions and the methodological/statistical questions were of interest. In this section, it is proposed to present the epidemiological context, questions and data for these two studies.

Studies II and IV were simulations studies. To simulate the data in Study II, realistic values were used for generating the cohort from which nested case-control studies were further sampled. These values came from an epidemiological study and the description of the data is part of this section. In addition to the simulation data, a real cohort was used to illustrate how the results applied in a real setting, both in Studies II and IV. The cohort used for this purpose is also presented in this section.

4.1 Collected real-world data

4.1.1 The CBC study (Study I)

4.1.1.1 Investigating risk factors for contralateral breast cancer

Contralateral breast cancer (CBC) is defined as a second primary breast cancer in the contralateral side, detected at least three months after the first breast malignancy.[71,72] Of all women with breast cancer, approximately 10 to 15% will develop an invasive contralateral breast cancer during the 20 years after initial diagnosis.[72] Risk factors for contralateral breast cancer have been investigated and studies have identified that ‘family history’ (i.e. up to third degree relatives who had a breast cancer),[73-76] a ‘non-ductal histological type’ of the initial breast cancer [71,77] and a young age at diagnosis of the initial breast cancer [76,78,79] are associated with an increased risk of developing contralateral breast cancer, while parity is often reported as a non-significant protective factor.[73,80,81] Multifocality of the initial breast cancer tumour has, to our knowledge, never been investigated. Since multifocality is reported to be associated with the higher-risk lobular histological type,[82,83] it could be a factor of interest to investigate. The epidemiological question which Study I aimed to answer was: Is multifocality of the initial breast cancer tumour a risk factor for contralateral breast cancer, and is parity confirmed as a protective factor for contralateral breast cancer?

4.1.1.2 The available data

4.1.1.2.1 Case data: contralateral breast cancer cases

Eight hundred and fifty three (853) patient cases of contralateral breast cancer, diagnosed between 1976 and 2005, were identified in the Stockholm-Gotland Cancer Register, which includes all patients diagnosed with cancer in this region since 1976. Patients’ medical charts were collected in order to retrieve additional variables of interest: known risk factors (‘family history’, ‘histological type’ and age at diagnosis date), potential confounders (chemo- and hormonal therapy) and the two potential risk factors to be investigated (multifocality of the breast tumour and parity).[72] All cases had their contralateral breast cancer three months or more after their first breast cancer and cases who had any other malignancy (except breast cancer) at any time before the contralateral breast cancer diagnosis date were excluded.

4.1.1.2.2 Potential control data: the 'Metastases study'

As data was collected for case patients within a well-defined cohort, the first idea about how to answer the epidemiological question was to collect controls in a nested case-control design, but it would be time and cost demanding.

It was thus suggested to reuse control data from another nested case-control study. A good candidate for that purpose was the 'Metastases study' which was a nested case-control study which also collected data for breast cancer patients registered in the Stockholm-Gotland Cancer Register. In this study, cases (191) were breast cancer patients who had metastases subsequent to breast cancer and controls (615) were sampled in a nested case-control design and remained free of metastases until their date of sampling. The same variables as for the contralateral breast cancer cases were retrieved from the medical charts as well as dates of diagnoses (including breast cancer, contralateral breast cancer and metastases). However, several additional inclusion/exclusion criteria had been used in the 'Metastases study': patients had to be younger than 76 years old at breast cancer diagnosis, had their breast cancer diagnosis in the restricted calendar period of 1997-2005 and had to be treated for their breast cancer with chemo- or hormonal therapy. An additional complexity was that the nested case-control sampling in this study was stratified on three factors: intended treatment (chemo-, hormonal therapy, or a combination of these), age category (<45, 45-54, >54 years), and period of breast cancer treatment (1997-2000 or 2001-2005).

Reusing this data set as the control data set for the contralateral breast cancer cases requires some attention, as the control data are a biased control sample. On the one hand, there is an overrepresentation of patients with metastasis, and on the other hand, as control patients were matched on time and on three other variables to the metastases cases, they cannot be readily reused to address a research question about a new outcome. The solution chosen was to have an IPW approach.

4.1.1.2.3 The underlying cohort

The condition enabling the use of these data sets to answer the epidemiological question using an IPW approach is to retrieve the correct study base, i.e. the underlying cohort from which the study patients (both cases and controls) were sampled. The different inclusion criteria for the two data sets described above, as well as the use of matching variables in the nested case-control sampling of the 'Metastases study', render the data situation complex. However as patients from both studies were included in the same register, and as dates of events (any malignancy and death) were available for all patients, both in the register data and in the two data sets, it was possible to align these data sets in order to use a weighted approach.

The data set retrieved from the Stockholm Breast Cancer Register comprised 32,153 breast cancer patients from 1976 to 2008 who were followed up for any other malignancy. The date of the patient's death, if this happened, was also available. This data set included patients with wider inclusion criteria than the two other data sets, which made it a worthy candidate to reconstruct an appropriate study base for the aligned set of case and control patients. An important part of the work, was to reconstruct a coherent study base in parallel to align and assemble the cases and controls data sets. This is represented in Figure 4.1.

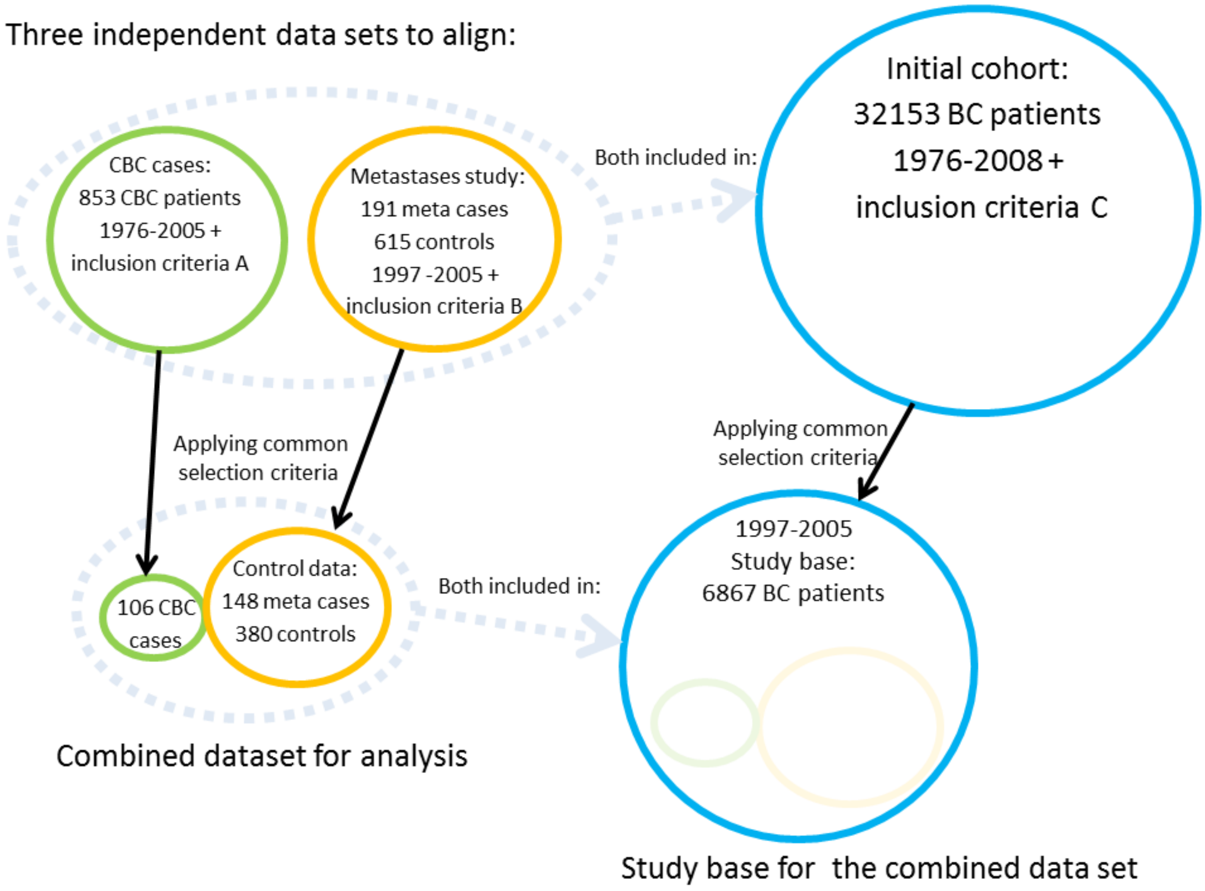

Figure 4.1: Alignment of the three data sets. As the inclusion criteria were different, they are called A (for the contralateral breast cancer cases), B (for the 'Metastases study') and C for the underlying cohort.

4.1.2 The SAMBAL study (Study III)

4.1.2.1 Investigating radiation therapy as a risk factor for subsequent lung cancer in female breast cancer patients

The effect of radiation therapy as a potential risk factor for subsequent cancer diagnosis has been investigated in several studies.[84,85] For women who have had postoperative breast cancer radiation treatment, an increased risk of lung cancer has been shown for at least 5 years, even decades in some cases, after the adjuvant treatment.[85,86] However, the main risk factor for developing lung cancer is smoking, and the risk of developing lung cancer after radiotherapy was shown to be particularly increased among smokers.[53,85,87] How these two carcinogens interact with each other is not fully understood, and the aim of the SAMBAL (i.e. molecular epidemiology of secondary lung cancer) study was to address the question: How does radiotherapy after breast cancer interact with smoking regarding the risk for lung cancer?

4.1.2.2 The nested case-control data

To address this question, participants were selected using a nested case-control design within the Swedish Cancer Register which includes all patients diagnosed with cancer in Sweden since 1958.[88] Cases were breast cancer patients, diagnosed between 1958 and the end of 2001 and subsequently diagnosed with lung cancer. Incidence density sampling was used to select matched controls (1 control per case) from breast cancer patients without any subsequent cancer diagnosis before their date of selection.

Seven hundred thirty (730) lung cancer cases and 726 controls were included in the study sample. As 5 years is estimated as a reasonable latency period for observing a radiation-induced solid tumour,[85,89,90] analysis was restricted to cases whose lung cancer occurred at least 5 years after the breast cancer diagnosis and their matched controls. Analysis was also restricted to patients who had information available on radiation therapy for breast cancer treatment, which led to a data set with 538 cases and 513 controls.

The data included the radiation dose received at each lung as well as the laterality of the breast and lung cancers, which means that each patient had a pair of dose measurements. In each of the case-control pairs of this nested case-control design, two levels of clustering were thus identified which will be of interest in the methodological approach of the present study.

In addition, the data were suspected to be overmatched as cases and controls were matched on decade of the breast cancer diagnosis and the treatment and radiation therapy protocols are strongly associated with time. The breast cancer patients entered the cohort from 1958 and at that time and until the mid-seventies, 80% of the patients received radiation therapy. As a consequence, patients in a matched set were more likely to share the same radiation therapy exposure, which in turn makes the set unused in any conditional analysis.

4.2 Study data in simulations studies

4.2.1 Simulation coefficients value in Study II

In the simulation settings in Study II, data was generated from a proportional hazards model using realistic values for the baseline hazard and the covariates regression coefficients. The chosen values of the variables were mimicking the observed values of a nested case-control study of coronary heart disease conducted within the Singapore Chinese Health Study.[91] The variance-covariance matrix and the means of the observed variables were used to first generate a multivariate normal distribution for gender, age, cholesterol, High-Density Lipoprotein (HDL), Systolic Blood Pressure (SBP), smoking status and antihypertensive treatment status, that were further rounded to the nearest integer (age), or dichotomised (gender, smoking status and antihypertensive treatment status). The correlation coefficients between the variables, together with the variables' mean value and variance are presented in Table 4.1.

Table 4.1: Variables means and variance-correlation matrix of Singapore Chinese Health Study nested case-control data

	Age (years)	Gender	Cholesterol (mg/dL)	HDL[a] (mg/dL)	SBP[b] (mm Hg)	Treatment status[c]	Smoking status
mean	63.1	0.51	203	53.8	136	0.27	0.18
Variance-correlation matrix							
Age	59.3	-0.158	0.049	0.148	0.267	0.130	-0.056
Gender		0.229	-0.201	-0.315	0.002	-0.001	0.273
Chol			1278	0.362	0.103	-0.096	-0.026
HDL				159	-0.054	-0.158	-0.130
SBP					488	0.187	0.039
Treatment						0.218	-0.128
Smoking status							0.190

[a] HDL: High-Density Lipoprotein;

[b] SBP: Systolic Blood Pressure;

[c] Treatment status: antihypertensive treatment status.

4.2.2 Data for illustration in Studies II and IV

In Studies II and IV, in addition to the simulation studies, how the methods work was illustrated in a real situation. In both studies, a real cohort of 75,856 brothers, sisters and children of all non-Hodgkin's lymphoma patients (probands) registered in the Swedish Cancer Register from 1958 to 2007 was used. In this cohort, the family members were followed on the age time-scale, from birth until emigration, death, or the end of the study (31 December 2007), whichever occurred first. The included variables were: gender of the family member, gender of the patient, type of relationship between the family member and the patient (sibling or child) and year of birth of the family member. The cohort was described and otherwise studied in detail by Lee et al.[92]

5 METHODS

The issues and limitations presented in the background section (Chapter 2) can be divided into two groups: issues due to the sampling design which would not arise within a cohort design (overmatching and subgroup analyses) and limitations due to the use of conditional logistic regression for analysing nested case-control data (reusing data, absolute risk estimation and analysis of clustered data).

To overcome the limitations of the conditional logistic regression in nested case-control data analysis, researchers have been developing novel statistical methods since the nineties. Their approaches can differ significantly, but a common point is that the matching between the cases and their controls is broken so that the controls are no longer tied to their matched case.[18,23,39,93-95]

Once the matching between cases and controls is broken, several authors have considered a weighted approach for the analysis in which the individuals are upweighted by their sampling probability.[23,39,93] This appealing approach, which originally comes from the survey literature, [36,37] and is also applied to case-control design [35] is used in this thesis as presented in this chapter.

5.1 Study base reconstruction and weighted partial likelihood

5.1.1 Study base reconstruction

The main idea when individuals from a sample are upweighted is to recover the study base, i.e. the underlying study population. Compared to the situation of Reilly et al. [35] with regular (matched) case-control studies, the particularity of the study base in a nested case-control study is the time aspect because the aim is to recover the cohort with the correct pattern of time at risk for the entire follow-up. This is illustrated in Figure 5.1, where Figure 2.3 is re-organised to present the nested case-control sample on the upper part of the graph and the remaining non-sampled individuals of the cohort on the lower part. In order to recover a valid representation of the study base by using the nested case-control data (in the dashed frame), only the non-cases need to be weighted as all cases from the cohort are usually part of the nested case-control sample.

Another question which needs to be answered before proceeding further is: Which study base needs to be reconstructed? Indeed, as presented in the former chapter, one may deal with two data sets which must be combined but have been sampled from two overlapping cohorts rather than from the same one. The challenge then is first to identify the correct study base which needs to be reconstructed.

A third aspect to be considered is the availability of information on the whole study base. As the idea is to up-weight individuals with the inverse of their sampling probabilities, information is needed on the exact sampling procedure, and on the study base from which the sampling was performed. This means that a well-defined cohort is needed for which basic information should be available, i.e. all relevant dates and variables used for the sampling. Once the study base has been identified and basic information is available, sampling probabilities and weights can be calculated.

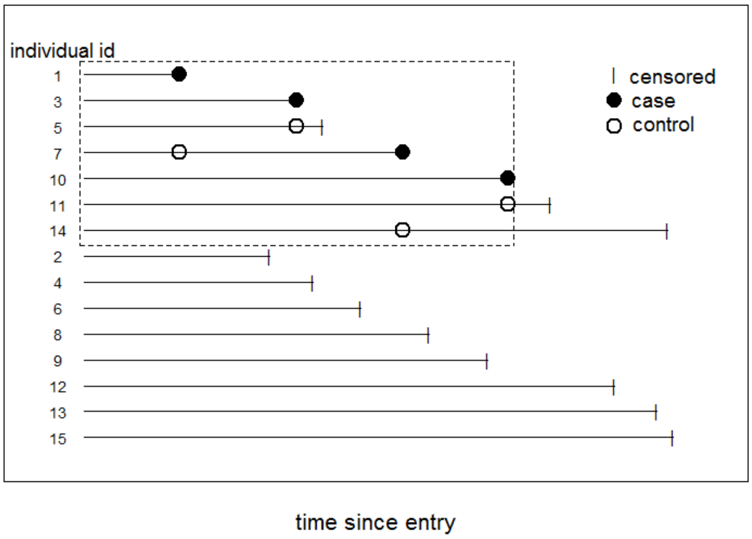

Figure 5.1: The nested case-control sample (in the dashed frame) versus the non-sampled individuals from the cohort in the lower part of the graph.

5.1.2 Kaplan-Meier weights

Samuelsen [39] suggested calculating the probability for an individual to be sampled in the nested case-control data with an expression which mimics the risk set sampling strategy. In a nested case-control design, all cases are usually included in the data, so that their sampling probability and hence their weight is equal to one. In contrast, the probability for an individual to be sampled as a control increases with the duration of his/her follow-up time and with the number of sampled controls at each event time, and decreases for an increasing number of individuals at risk at each event time (i.e. the size of the risk set R_i, which is itself evolving with time).

At each event time t_i, an individual k with starting time S_k ($S_k \leq t_i$) and censoring or event time T_k ($T_k \geq t_i$) is available to be sampled. Assuming m_i controls were sampled for case i from a risk set of R_i individuals, the probability p_{ki} that k is sampled at time t_i for case i is given by $m_i/(R_i$ - 1). Multiplying the probabilities of not being sampled (1 - p_{ki}) during the entire follow-up (i.e. between S_k and T_k), and subtracting this product from one, gives the probability p_k for individual k to be sampled at least once during his/her follow-up time. In short [39]:

$$p_k = 1 - \prod_{i,\, S_k \leq t_i \leq T_k} \left[1 - \frac{m_i}{R_i - 1}\right] [1 - Y_k(t_i)] \qquad (5.1)$$

where $Y_k(t_i)$ is an indicator of the case-control status of individual k at time t_i.

The weight w_k is the inverse of this probability. The weights aim to 'reconstruct' the number of individuals at risk in the full cohort, idea which is similar to that which is done in survey inference.[37]

In our small cohort in Figure 5.1, applying Equation 5.1, the weights are 6.7 for individual 5, and 2.5 for individuals 11 and 14. All other individuals are cases and get a weight of 1. Exploring how these weights reconstruct the pattern of time at risk in the cohort gives the following results: At time t_1, where 15 individuals were included in the risk set, Samuelsen weights recover 16 (15.7, rounded) individuals, and at time t_2, t_3 and t_4, the weights recover 15, 7 and 6 individuals for numbers that were actually 13, 9 and 6, respectively. It can be noted that this weighting system appears to be effective, even with such a small cohort.

On a practical level, the weights can be calculated by using simple readily available software commands: a Kaplan-Meier analysis of the cohort provides the number of individuals in the risk sets at each event time. From these numbers, with simple algebra only including subtraction, division and cumulative products, the probabilities and thus the weights are computed. As the Kaplan-Meier analysis is the central tool used for calculating these weights, this name is usually given to the weights developed by Samuelsen.[96,97]

5.1.2.1 Stratified Kaplan-Meier weights

In case of stratified sampling, i.e. sampling involving additional matching on a confounder Z, the expression 5.1 can be generalized [38,98]:

$$p_k = 1 - \prod_{i,\, S_k \le t_i \le T_k} \left[1 - \frac{m_i}{R_i^Z - 1}\, I(z_k = z_i) \right] [1 - Y_k(t_i)] \qquad (5.2)$$

where R_i^Z is the number of individuals at risk at t_i who have the same value for the confounder as the case and $I(z_k = z_i)$ is the indicator that case i and individual k have the same value for the confounder Z.

Regarding the implication in the computing aspect, the only difference is that Equation 5.1 has to be applied in each of the strata defined by the matching factor. However, in cases of fine matching, the Kaplan-Meier weights could fail to provide a correct representation of the cohort, as high sampling probabilities could lead to weights which are too light.[97]

5.1.3 glm/gam weights

Another type of weights which is also easy to implement was proposed by Kim and De Gruttola.[93] The weights are obtained by modelling the sampling probability of the controls with a parametric model (such as logistic regression (glm) or generalised additive models (gam)) which includes the available covariates which drove the sampling and the time span during which the individual was available for sampling. The model is run on the whole cohort from which the cases are excluded. As the time is included in the model, the increasing probability of selection with time at risk is accommodated by the model. These weights are referred to as the glm/gam weights.[96,97,99]

The calculation of the glm/gam weights requires all cohort members to have a variable which indicates whether they have been sampled for the nested case-control study. In contrast, the Kaplan–Meier type of weights does not need this sampling indicator to calculate the risk set sizes involved in the weights calculation. This is an advantage when working with data which were anonymised to ensure individuals' data protection after the sampling has been performed within the study base. Another feature of the glm/gam weights is that they require some modelling which is not needed with the Kaplan-Meier method.

5.1.4 Weighted partial likelihood statistical method

Analysing weighted data in order to obtain the coefficients estimate β is done by maximising a weighted partial likelihood (pseudo-likelihood) whose expression is [39]:

$$L(\beta, \gamma) = \prod_{t_i} \frac{\exp[\beta' X_i + \gamma' Z_i\,]}{\sum_{k \epsilon R*_i} \exp[\beta' X_k +\, \gamma' Z_k]\,.\,w_k} \qquad (5.3)$$

where $R*_i$ is the collection of all cases and sampled controls at risk at time t_i, and w_k is the weight for individual k. Any individual in the pooled data is a subject who is followed from starting time until the outcome or censoring date. Figure 5.2 represents risk sets $R*_i$ which include all individuals who are pooled together, weighted and followed up.

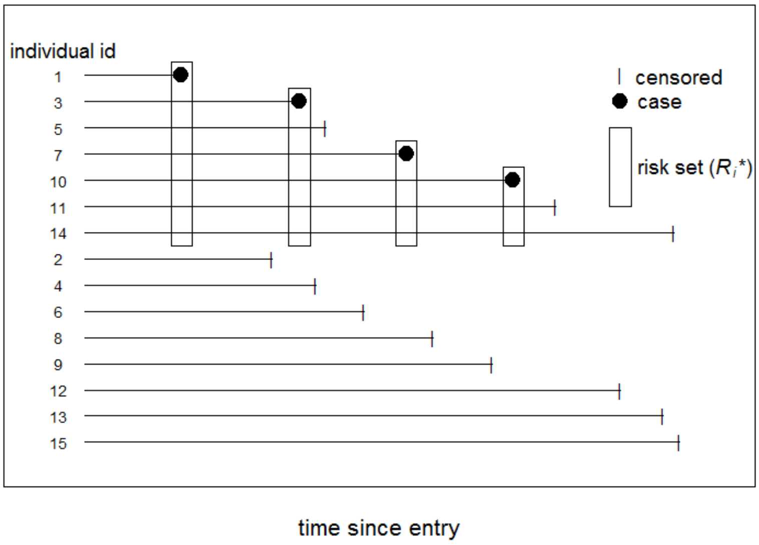

Figure 5.2: Pooled individuals of the nested case-control sample and risk sets R_i*. The data are re-organised as in Figure 5.1. The three sampled controls represent the non-cases in the whole cohort and are weighted.

In practice, the analysis is performed in statistical software by using a weighted Cox regression with the pooled data and unique individuals who are characterised by a time-to-event or time-to-censoring, as in the (usual) Cox analysis of an entire cohort.[100] Samuelsen derived an asymptotic variance for β that accounts for the Kaplan-Meier weights, [39] and Kim, [101,102] using the same weights proposed an approximate jackknife standard error. However, using a robust variance estimator is usually considered as a reasonable choice for the weighted analysis.[96,97]

In this approach, as all cases and sampled controls are pooled together, an initial benefit appears: while each contribution in the likelihood (2.8) only included the case and the m_i sampled controls, the likelihood (5.1) uses all cases and sampled controls who are still at risk. By using more individuals in each contribution of the likelihood, a gain in statistical power is expected compared to the conditional logistic regression analysis.

5.1.5 Achievements with the weighted approach

5.1.5.1 Hazard ratio estimates

The way in which the weighted method performs in data analyses has been studied through simulation studies.[18,93,96,97,99-103] Whatever the type of weights used, the weighted method provides valid estimates of the hazard ratio β with a slightly better precision than the conditional logistic regression analysis. Comparing the use of Kaplan-Meier and glm/gam weights, they give comparable estimates and show similar empirical variances,[96,97,99] but when censoring time is linearly correlated with a covariate, the glm/gam weights method is slightly more efficient than the Kaplan-Meier method,[99] and in case of fine matching, the glm/gam weights perform better than the Kaplan-Meier weights.[97]

5.1.5.1.1 Hazard ratio estimates in case of stratification

Considering the expression 2.5 (i.e. $h_i\ (t|X_i,\ Z_i,\ \beta,\ \gamma) = h_0(t)\ exp(\beta' X_i + \gamma' Z_{i,}))$, an important achievement of the weighted method is the ability to estimate regression coefficients for covariates which were matching factors, i.e. it enables estimating the γ's.

As the matching is broken, the estimation of the regression coefficients are no longer performed within the risk sets R'_i as in the conditional logistic regression but on pooled data, i.e. the risk sets R^*_i. As the cohort is reconstructed thanks to the weights, the γ's can be estimated in the same way as can be done in a cohort analysis, provided the weights have been correctly calculated using the stratified expression 5.2.

If estimating the β's while analysing a stratified nested case-control sample is the only area of interest, Støer et al.[97] showed that the accuracy of the estimates of the exposure's hazard ratio was similar for unstratified weights (Equation 5.1) and stratified weights (Equation 5.2) provided one adjusted in the analysis for the confounder(s) which were used as matching variable(s). If this unexpected result could be demonstrated to be true in general, it would represent a useful simplification when using the method and would also mean that, in the case of fine matching, the reconstruction of fine strata is unnecessary, avoiding a problem which arises with these weights, as mentioned above.[97]

5.1.5.2 Absolute risk estimation with weighted partial likelihood

The ability to validly estimate both β's and γ's gives the opportunity to revisit the Breslow estimator and develop an expression which exploits the weighted approach. It seems that this idea was being investigated in 2012, as just previous to the approach outlined in Study II, Cai et al. [104] developed an estimator for prognostic accuracy of biomarkers from nested case-control data using the weighted approach of Samuelsen to estimate the coefficients of the risk factors (β) and adapting the Breslow estimator to the weighted situation. The adapted cumulative baseline hazard Breslow estimator is:

$$H_0(t) = \sum_i \frac{I\,(t_i \leq t)}{\sum_{k \in R*_i} \exp[\beta' X_k + \gamma' Z_k]\, w_k} \qquad (5.4)$$

with $R*_i$, the collection of all cases and sampled controls at risk at time t_i, and w_k the weight of control k, as in the likelihood (5.3). This development provides a new way to consider absolute risk estimation with nested case-control data.

5.2 Other likelihood approaches

In parallel with the approaches using a weighted partial likelihood, other methods have been developed. Scheike and Juul [94] and Saarela et al. [18,103] suggested a complete cohort likelihood approach in which the cohort sampling design is treated as a missing data problem, and which requires modelling of the distribution of the partially observed covariates. Another approach was explored by Keogh and White [95] in which multiple imputation techniques using the fully observed data to fit the imputation models are used, addressing thus the nested case control data analysis as a missing data problem as well. All these methods have been investigated and compared in several studies, and all of them present advantages and disadvantages.[18,96,103,105]

The approaches just mentioned above, as well as the weighted approach, require data to be available on basic information regarding time-to-event(s) or censoring for all participants of a well-defined cohort. In cases of stratified sampling the cohort should also contain information on the matching variables for all of its members. The full likelihood and multiple imputation methods will benefit from having access to more information in the cohort, should this information be available.[105] Other variables than the times to event(s) or censoring and matching variables which would be available in the full cohort are not used to compute the weights in the weighted approach (with Kaplan-Meier weights), but can be used to improve the imputation model or the model that describes the covariates distribution in the full likelihood method. As both methods also require some modelling, they are vulnerable to model misspecification and they also require more programming.[18,96,103] In addition, these approaches need an indicator telling which individuals were sampled and which were not, among all cohort's members, an information which is not always available.

Last but not least, as the nested case-control design is sometimes chosen in order to reduce the computing burden, when for example, mega data-bases involving large cohorts with multiple time-varying exposures or covariates are handled [106], the methods involving full likelihood or multiple imputation would definitely show some drawback in such cases.

5.3 Simulation methods

Simulating is an effective tool to explore a hypothesis or asses the performance of statistical methods in relation to the known truth. Simulation methods were the main tool used in Studies II and IV. Following the guidelines outlined by Burton et al.,[107] the simulation settings were decided according to the aims of the simulation. Cohorts were simulated whose estimates served as reference for all subsequent analyses of the nested case-control data which were further sampled in these cohorts and analysed with various methods (i.e. we adopted what Burton et al, called 'moderately independent simulations' [107]). For the different scenarios (i.e. different types of cohorts), however, fully independent simulated data were generated.

5.3.1 Generating the data set

In order to generate the data set, the use of realistic scenarios was aimed for reflecting plausible epidemiological situations; however, settings were also challenged with more extreme (unrealistic) values.

5.3.1.1 Time-to-event

The distribution of the time-to-event was generated using the method described by Bender et al. and Crowther et al. [108,109]:

Simulating time-to-event starts from Equation (3.2) i.e. $F(t)$ = 1- exp (-$H_0(t)$ exp ($\beta'X + \gamma'Z$)). As the values taken by this distribution are included in the interval [0, 1], as well as the values taken by (1 – F), simulated values can be generated from a uniform distribution (U [0, 1]). Using U = exp (-$H_0(T)$ exp ($\beta'X + \gamma'Z$)), (with T the time-to-event variable), time-to-event values can be generated, provided H_0 can be inverted: $T = H_0^{-1}$[-log(U) exp(-$\beta'x$)].[108,109] When T cannot be solved analytically, iterative root finding methods are used.

5.3.1.2 X's and Z's distribution

In Study II, the data used were generated from the covariates mean and the variance-covariance matrix retrieved from a study of coronary heart disease nested in the Singapore Chinese Health Study.[91] The correlation structure between the variables from one scenario to another was not modified.

In contrast, in Study IV, data were generated from normal and binomial distributions, using mean values which were retrieved or inspired by the epidemiological data analysed in Study III, but in the different scenarios explored, some variation was made in the way the Z variables were correlated to the exposure variable. This way of generating data gave a huge freedom to simulate variables which were defined according to their role relative to the exposure. Beside the exposure, variables were generated which were either independent risk factors, confounders or effect modifiers. Table 5.1 describes the variables used in the main simulation scenario of Study IV.

5.3.1.3 Vectors of coefficients β and γ

The same differences between Studies II and IV apply for the vector of coefficients *β* and *γ*. In study II, the values were retrieved from the same, already mentioned, study of coronary heart disease. In some scenarios though, these values were modified. In Study IV, the coefficients were inspired by the epidemiological data of Study III, but different values were explored for the β's and γ's in the scenarios which were considered. The chosen values for the β's and γ's in our main scenario are presented in Table 5.1.

5.3.1.4 Baseline hazard

In Study II, the baseline hazard which best described the time-to-event distribution in the study of coronary heart disease was found to be a Weibull distribution and this was used in the final setting. In Study IV, a constant baseline hazard was used in the main setting (Table 5.1) and a Weibull baseline hazard in additional settings.

Table 5.1: Variables distribution and parameters' value in the main simulation setting of Study IV

Variable	Distribution	Coefficient	Hazard ratio
Exposure X_e	Binomial (N, p*)	$\beta_e = 0.405$	1.5
Independent risk factor X_{irf}	Normal ($\mu = 0$, $\sigma = 10$)	$\beta_{irf} = 0.020$	1.02
Confounder X_c	Binomial (N, 0.5)	$\beta_c = 0.953$	2.6
Effect modifier X_{em}	Binomial (N, 0.3)	$\beta_{em} = 1.386$	4
Interaction coefficient between X_e and X_{em}: $\beta_{interact}$		$\beta_{interact} = 0.69$	2
*p is defined by the association between X_e and X_c:	$P(X_e \mid X_c) = \exp(\frac{\log(4)+\log(1/4)*Xc}{1+\exp(\log(4)+\log(1/4)*Xc})$		
Baseline hazard	Constant	$h_0 = 0.0005$	

5.3.1.5 Censoring time

In both Studies II and IV, the censoring time was generated with an exponential function, and a maximum follow-up period length was also included. In addition, in some of the explored scenarios in Study IV, censoring distribution was made depending on the effect modifier.

5.3.2 Sampling of nested case-control studies in the simulated cohorts

Once the cohort data were set up, nested case-control sampling was simulated in Studies II and IV. For each scenario, 500 cohorts of N individuals (N varying between 50,000 and 100,000) were simulated, and in each of them, one nested case-control study was sampled before various analyses were performed on the nested case-control sample and the cohort. The nested case-control sampling involved the choice of the number of controls per case, and the choice of whether or not to perform a stratified sampling. Different scenarios corresponded to different choices. The number of controls per case was either two or five, and the sampling did not involve matching for several scenarios, while for other scenarios, matching was performed on one or two confounders. Fine matching was also performed.

5.3.3 Assessment of the performance of the statistical methods

In both studies, the performance of the statistical methods used to analyse the nested case-control data set in comparison to the estimates obtained from the cohort was assessed. The bias and the accuracy of the estimates were evaluated.

5.4 Statistical software

All simulation studies and analyses of the generated data and the real data sets were performed in R. Codes used in Study II were developed by Salim and collaborators. Unique codes were developed to generate the data in Study IV. To perform the nested case-control sampling within the created cohort, codes were used which had already been developed within the team. Unique codes were also created to calculate the Kaplan-Meier weights which form the central part of the weighted partial likelihood method presented above. Part of this development is available at: http://www.meb.ki.se/~biostat/. These pieces of codes show how a cohort can be generated, and a nested case-control study be further sampled, how weights can be calculated, and how they can be assigned to the controls. The *R* packages which are needed to calculate the Kaplan-Meier weights are the *survival* and *plyr* packages.

Støer et al. developed the *multipleNCC* package, implemented in *R* and available on CRAN since autumn 2014.[110,111] This package fits Cox proportional hazard models with a weighted partial likelihood, including different types of weights (among which the Kaplan-Meier and glm/gam weights), and accommodates the reuse of controls for several endpoints. However, the package requires all data to be in a single data set, which is not always possible in real research situations, such as in Studies I and III.

6 OVERVIEW OF THE FOUR STUDIES (RATIONALE AND OBJECTIVES)

Each of the issues presented in Chapter 2 was tackled in the thesis. Reusing nested case-control data to address a research question regarding another outcome (first issue) was the central point of interest in Study I. However, as breaking the matching (which was done in all four studies) is similar to reusing cases and controls [39], the theme 'reusing the data' was somehow explored in all studies.

The absolute risk estimation (second issue) was studied in detail in Study II with simulation studies, and performed in Study III on a real-world data set. Breaking the matching to overcome problems due to overmatching (third issue) and clustering (fourth issue) was addressed in Study III, and subgroup analyses (fifth issue) were investigated in Study IV by means of simulation studies.

In this latter still unpublished study, we also investigated in detail how well the Kaplan-Meier weights were able to reconstruct the number of individuals at risk over time in the cohort and in subgroups. Study IV shed new light on the results which were previously obtained in the first three studies.

As previously stated, in Studies I and III, we worked with real data sets, to investigate epidemiological questions, in addition to the methodological/statistical questions which were the main interest. Studies II and IV were mainly simulations studies where the real data set analysis served as illustration.

This chapter aims to present the rationale and the specific objectives of each study. As several research questions were addressed in several studies, Table 6.1, at the end of this chapter, summarises questions and studies in which they were explored.

6.1 Study I: Reusing nested case-control data to address a research question regarding a new outcome

6.1.1 Why a methodological question was addressed in Study I

While Study I was motivated by the epidemiological question regarding the risk factors for contralateral breast cancer, we focused our interest on the methodological and statistical questions. The reason lies below.

All published studies using the IPW approach to reuse nested case-control data aimed to validate the method and used various approaches to compare the estimates.[38,96,97,100,112] The authors had either access to the full cohort from which the nested case-control was sampled (simulation studies), [96,97,100] or controls sampled for the new outcome which could be used in the validation step.[38,100,112] The advantages of the method were demonstrated in these ideal illustrative data analyses. All reported an increased efficiency of the nested case-control design when additional controls are used which were previously sampled to study another outcome in the same underlying cohort. However, the method has been largely ignored by medical researchers and has not found its way into routine use by biostatisticians and epidemiologists.[29,101] To the best of my knowledge, there is no published work of an application of the method to a setting where the data have features typical of real research situation, such as cases and available controls being sampled from different, but overlapping, cohorts and with different inclusion and sampling criteria.

In this application, beyond the work of Salim et al. [38,100] and Støer et al. [96,97,112] we wanted to assess the advantages and limitations of reusing nested case-control data in a real situation where the second study did not gather any controls, so that the only option would be to reuse prior control data. We wanted also to clarify the conclusion of Støer and Samuelsen [96,97] concerning the choice of weights, highlighting how the feasibility of calculating several types of weights may depend on the data at hand.

6.1.2 The methodological questions

To assess the advantages and limitations of reusing nested case-control data in a real and complex situation includes assessing different items which are listed below:

- What is the impact of using an appropriate method rather than a naïve statistical analysis when reusing nested case-control data?
- Is there a difference in the estimates when reusing a whole nested case-control data set or when only the sampled controls of this data set are reused?
- What is the impact of the accuracy of the reconstruction of the underlying cohort from which the weights are calculated?
- What is the impact of using unstratified weights instead of stratified weights to estimate covariates coefficients in a matched design?
- What is the impact of the choice of data which are reused to analyse the new outcome (when a choice is possible) on the final data set and for the interpretation of the estimates?
- What is the impact of using selection dates of the controls when the censoring date in the nested case-control data set is ignored?
- Is there a practical advantage to using Kaplan-Meier or glm/gam weights?

6.2 Study II: Estimating absolute risk with nested case-control data

6.2.1 Why Study II was performed

Langholz and Borgan [45] developed an estimator for the cumulative baseline hazard (Chapter 2). However, it remained unclear how to accommodate the approach to estimate absolute risk with nested case-control data in case of stratified sampling, because the conditional logistic regression which is used to retrieve the estimates will not be able to estimate the coefficients for the matching factors, while they are needed in the absolute risk estimation.[50]

The weighted method developed by Cai et al. [104] (Chapter 4) for estimating absolute risk with nested case-control data was used by Zhou et al. [113] who analysed data collected in a nested case-control design to estimate the absolute risk of developing rheumatoid arthritis. Although they reported results obtained with the weighted method, the article did not provide details on how they implemented the method.

While both the Langholz-Borgan and the weighted methods are not new, the implementation of these methods using standard statistical software is still lacking, as well as a complete comparison of the absolute risk estimations with the two methods, especially in case of stratification.

6.2.2 Research questions

The study thus aimed to compare the two methods for estimating absolute risk from nested case-control data, in both simulations studies and real-world data, to examine their performance and discuss the relative merits of each. Specifically, the questions which the study aimed to answer were:

- Are these methods suitable to estimate absolute risk from matched nested case-control data?
- Are the Langholz-Borgan and weighted methods comparable in terms of accuracy and efficiency?

6.3 Study III: Advantages of weighted partial likelihood over conditional logistic regression in analysing clustered and overmatched nested case-control data

6.3.1 Why a methodological question was addressed in Study III

While addressing an epidemiological question regarding how smoking and radiation therapy interact in the risk of developing lung cancer in female breast cancer patients, we however focused our interest on the methodological and statistical aspects. The reason lies below.

The collected data included information on paired organs: radiation doses assessment for both lungs and laterality of the cancers (breast and lung). The data were gathered in a nested case-control design which was also overmatched, due to the matching on breast cancer decade of diagnosis (Chapter 4). The overmatching and clustering of the data rendered the weighted partial likelihood approach attractive compared to conditional logistic regression analysis. Furthermore, as the sample was performed within the Swedish Cancer Register, the weighted partial likelihood method could be used.

As overmatching problems result in losing matched sets in the analysis (the sets where case and controls share the same exposure value), breaking the matching was thought to be a potential solution worth exploring. The idea was inspired by Samuelsen [99] who suggested pooling nested case-control data to reduce the efficiency loss due to missing covariates (another unrelated topic but with some similar consequences to overmatching) and by

Brookmeyer [58] who, for case-control studies, showed that pooling the data from several strata could help solving or mitigating a problem due to overmatching.

Regarding the clustering feature, the weighted approach will allow reducing the two levels of clustering (set and individual levels) to a single level (individual level) and when using robust standard error in the analysis, the clustering at the individual level will be accounted for.

In addition to mitigating this problem of overmatching and handling the paired data, the weighted partial likelihood will enable the estimation of cumulative risk studied in detail in Study II, which was considered a further advantage.

The specific statistical question which the study aimed to answer was:

- What are the advantages of using weighted Cox regression compared to conditional logistic regression in analysing overmatched and clustered nested case-control data?

6.4 Study IV: Analysing subgroups with nested case-control data

6.4.1 Why Study IV was performed

As subgroup analyses are regularly performed in nested case-control studies [61-64] and as it remains unclear if these analyses can provide valid estimates, simulation studies were performed, which aimed to answer the following specific questions:

- Are subgroups analyses with nested case-control data valid when a subgroup is defined by a covariate measured at baseline?
- Is one method to be preferred among weighted likelihood method and conditional logistic regression for subgroup analyses of nested case-control data?
- How well does the weighting system reconstruct the cohort and each subgroup?
- How well does the weighted likelihood method with unstratified weights in recovering the exposure coefficient in each subgroup in case of matched design?

Table 6.1: Specific methodological research questions and studies where these were addressed.

Theme, context and questions:		Addressed in
Reusing data: Breaking the matching to address a research question regarding a new outcome		
Question 1	What is the impact of the accuracy of the reconstruction of the underlying cohort from which the weights are calculated?	Study I
Question 2	Is there a difference in the estimates when reusing a whole nested case-control data set or when only the sampled controls of this data set are reused?	Study I
Question 3	What is the impact of the choice of data which are reused to analyse the new outcome (when a choice is possible) on the final data set and for the interpretation of the estimates?	Study I
Breaking the matching and weighting individuals		
Question 4	How well does the weighting system reconstruct the cohort and each subgroup of the cohort?	Studies III, IV
Question 5	What is the impact of using unstratified weights instead of stratified weights to estimate covariates coefficients in matched designs?	Studies I, III, IV
Question 6	What is the impact of using an appropriate method rather than a naïve statistical analysis when reusing nested case-control data?	Study I
Question 7	What is the impact of using selection dates of the controls when the censoring date in the nested case-control data set is ignored?	Study I
Question 8	Is there a practical advantage to using Kaplan-Meier or glm/gam weights?	Studies I, III
Estimating absolute risk		
Question 9	Are the Langholz-Borgan and weighted methods suitable for estimating absolute risk from matched nested case-control data?	Study II
Question 10	Are the Langholz-Borgan and weighted methods comparable in terms of accuracy and efficiency?	Study II
Overmatching		
Question 11	Does weighted partial likelihood help overcome overmatching?	Study III
Clustered data		
Question 12	What are the advantages of using a weighted Cox regression with clustered data?	Study III
Subgroup analyses		
Question 13	Are subgroups analyses of nested case-control data valid when a subgroup is defined by a covariate measured at baseline?	Study IV
Question 14	Is one method to be preferred among weighted Cox regression and conditional logistic regression for subgroup analyses of nested case-control data?	Study IV

7 RESULTS AND DISCUSSION

In this thesis, both epidemiological and methodological/statistical questions were addressed, but the choice was made to focus on the latter type of questions. In addition, as summarised in Table 6.1, several questions were explored in different studies. The structure of Chapter 7 will reflect these features.

The results for all methodological/statistical questions are first considered where the same order as in Table 6.1 is followed. The epidemiological results (Studies I and III only) are presented in the last part of the chapter.

7.1 Results regarding the methodological/statistical questions

7.1.1 Reusing nested case-control data to address a research question regarding a new outcome

For this theme (reusing nested case-control data) and context (answering a question regarding a new outcome), Study I gave the opportunity to answer several specific questions. Reusing nested case-control data which were sampled from an overlapping cohort to study another outcome, risk factors were investigated for contralateral breast cancer (our new outcome of interest).

Table 7.1 presents the estimates obtained in several analyses. In the first column of the table are the results from our published article.[114] This main analysis uses **cases and controls** from the reused data set as control data, together with the contralateral breast cancer cases; this combined data set is then analysed with a **weighted** Cox regression (i.e. weighted likelihood method). The weights are **stratified** on the matching factors used to sample the controls and calculated with a Kaplan-Meier analysis of the **appropriate study base** for the current study. The **covariates** which are included in the model are the four risk factors (multifocality of the breast cancer tumour, parity, histological type of the tumour and 'family history'), in addition to the potential confounders (adjuvant treatment and age). All terms highlighted in bold style in the text above are features of the main analysis. All subsequent analyses involve a change regarding these features.

The other columns of Table 7.1 present the results from analyses which were performed by making some variation of the first analysis. In the text which follows, the terms which are written in bold emphasise the feature under interest, i.e. the element which is modified compared to the first analysis.

The estimates in column 2 were obtained when adding a **covariate** (the size of the breast cancer tumour). This variable could have indirectly influenced the sampling of the patients in the 'Metastases study' (and hence the current analysis) as this study included patients who had freshly frozen tumour material available for genetic analysis, which could be related to the size of the tumour. The analyses reported in all subsequent columns of Table 7.1 also included this covariate. The estimates in column 3 were obtained with weights which were calculated from a Kaplan-Meier analysis of the **study base relevant to the 'Metastases study'** which included more patients than the study bas used in the main analysis. To obtain the estimates in column 4, only the **sampled controls** from the 'Metastases study' were reused to prepare the control data, instead of both metastases cases and controls. In column 5, the results are given where **unstratified weights** were used in the weighted Cox regression analysis, and in column 6, a **naïve unweighted** analysis was performed. Each of the columns 3 to 6 will be compared with column 2, while the latter will be compared to the first column.

7.1.1.1 Question 1

> What is the impact of the accuracy of the reconstruction of the underlying cohort from which the weights are calculated? (Study I)

In parallel with the correct study base (relevant to the current study), the study base relevant to the 'Metastases study' was also reconstructed, and the weights with both study bases were computed. The study base relevant to the 'Metastases study' was not completely irrelevant to the current study, as most of the criteria to align the data sets were imposed by the 'Metastases study'. This latter study base was larger than the appropriate one as it included patients with several malignancies, representing around 13% of the patients. The estimates were, however, very similar, when using either study base to calculate the weights (Table 7.1, column 2 and 3). This can be expected when a large study base is used (both study bases were large), so that the risk sets were large at any event time and the impact on the weights is limited as shown in Figure 7.1. That could, however, be problematic for small study bases.

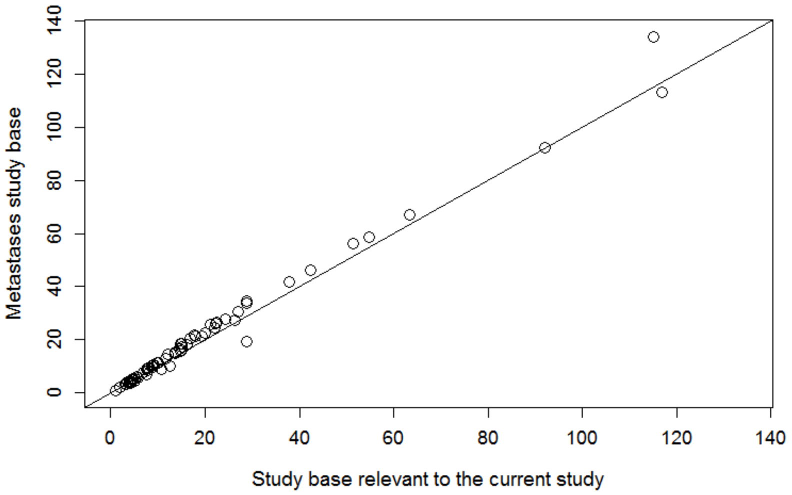

Figure 7.1: Correlation between the stratified weights computed with the study base relevant to the current study and the study base relevant to the 'Metastases study'.

Table 7.1: Adjusted risk estimates: hazard ratios and (95% confidence intervals) from weighted Cox regression analyses (five first columns) and unweighted Cox regression analysis (last column).

Risk factors	Main analysis[a]	Additional covariate[b]	Study base for meta[c]	Sampled controls[d]	Unstratified weights[e]	Unweighted[f]
Non-multifocal tumour (ref.)	1	1	1	1	1	1
Multifocal tumour	1.99 (1.07, 3.70)	1.94 (1.03, 3.68)	1.94 (1.02, 3.69)	1.97 (1.03, 3.77)	1.91 (1.04, 3.51)	1.56 (1.03, 2.39)
Nulliparous (reference)	1	1	1	1	1	1
Parity	0.40 (0.18, 0.89)	0.42 (0.18, 0.95)	0.41 (0.18, 0.92)	0.39 (0.17, 0.91)	0.50 (0.24, 1.02)	0.82 (0.48, 1.39)
Ductal histological type (ref.)	1	1	1	1	1	1
Non-ductal histological type	2.09 (1.21, 3.59)	2.11 (1.20, 3.70)	2.14 (1.22, 3.77)	2.14 (1.21, 3.77)	2.10 (1.24, 3.55)	1.79 (1.23, 2.60)
No family history (reference)	1	1	1	1	1	1
Positive family history	1.91 (1.11, 3.28)	2.04 (1.17, 3.55)	2.06 (1.18, 3.59)	2.06 (1.18, 3.62)	2.13 (1.29, 3.54)	1.59 (1.11, 2.27)
Chemotherapy (reference)	1	1	1	1	1	1
Hormonal therapy	0.71 (0.39, 1.26)	0.70 (0.37, 1.31)	0.64 (0.34, 1.19)	0.72 (0.38, 1.37)	2.81 (1.55, 5.11)	1.95 (1.17, 3.26)
Chemo + Hormonal therapy	0.57 (0.30, 1.07)	0.59 (0.30, 1.17)	0.59 (0.30, 1.17)	0.60 (0.30, 1.20)	0.82 (0.42, 1.62)	0.77 (0.43, 1.38)
Age <45 (reference)	1	1	1	1	1	1
Age 45-54	1.23 (0.62, 2.44)	1.29 (0.63, 2.64)	1.22 (0.60, 2.51)	1.31 (0.63, 2.69)	2.07 (1.063, 4.02)	1.69 (0.98, 2.93)
Age >54	0.96 (0.49, 1.88)	0.97 (0.48, 1.96)	0.95 (0.47, 1.89)	0.96 (0.47, 1.93)	1.09 (0.55, 2.14)	1.18 (0.69, 2.03)
Tumour size (mm)	--	0.99 (0.96, 1.02)	0.99 (0.96, 1.02)	0.99 (0.96, 1.02)	0.99 (0.96, 1.02)	0.99 (0.97, 1.02)

[a] Results obtained from a weighted Cox analysis, adjusted for confounders, with weights computed by stratified Kaplan–Meier analysis of the relevant study base.

[b] Same as the main analysis when adding the tumour size of the breast cancer as an additional covariate.

[c] Same as the analysis in the 2nd column with weights computed with the imperfectly reconstructed study base, i.e. study base relevant to the 'Metastases study'.

[d] Same as the analysis in the 2nd column with the 398 sampled controls only instead of the whole 'Metastases study' case-control dataset

[e] Same as the analysis in the 2nd column with unstratified weights.

[f] Naïve unweighted analysis adjusted for assumed confounders.

Relating to the question of the study base reconstruction, there were also some concerns about the role of the tumour size in the sampling of the ‘Metastases study’. It was not possible to include this variable in the weights, because the sampling was not stratified on this variable, but the variable was included in the analysis to take into account the possible bias due to this variable (Table 7.1, column 2). In the published paper, this variable was removed as it did not have any impact on the results (Table 7.1, column 1).[114]

7.1.1.2 Question 2

> Is there a difference in the estimates when reusing a whole nested case-control data set or when only the sampled controls of this data set are reused? (Study I)

The results presented in table 7.1, in the second and fourth columns, show that the estimates were similar when reusing the 528 patients (148 metastasis cases and 380 controls of the ‘Metastases study’) or when reusing only the 398 sampled controls of the ‘Metastases study’ among whom 18 patients developed a metastasis after being sampled (i.e. they became cases later during their follow up time).

This is not surprising. The 398 patients, including the 18, were weighted, and with their weights, these patients were representative of the study base. In particular, the 18 were representative of the special subgroup of patients who developed metastases and became cases during their follow-ups (i.e. the 148 case patients in the ‘Metastases study’). Figure 7.2 compares the actual number of individuals in this subgroup who remain in the nested case-control data set over time and the corresponding number recovered with the weights when retaining only the patients who had been sampled as controls and who are weighted.

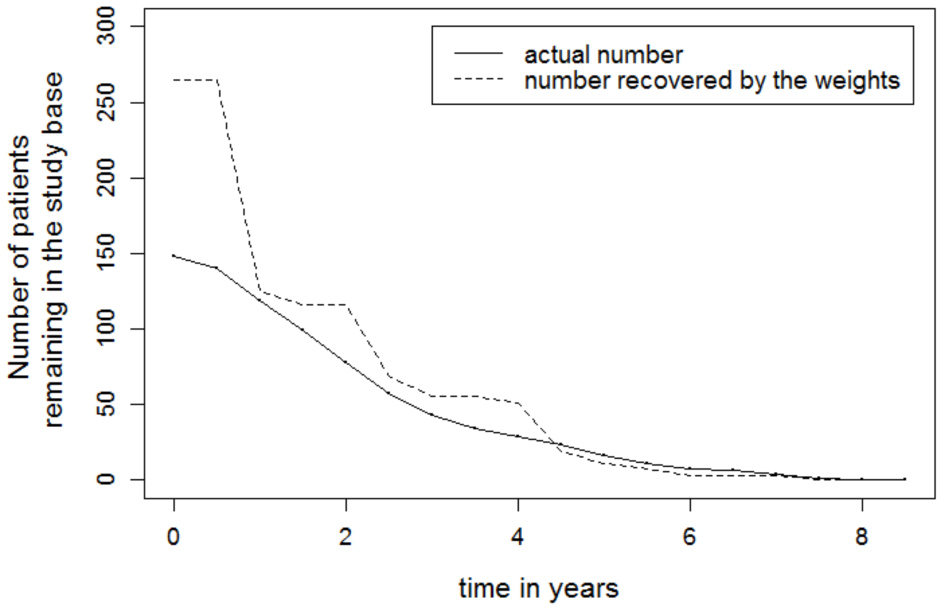

Figure 7.2: Number of case patients in the ‘Metastases study’ over time and corresponding numbers recovered with the weights.

For example, when the follow-up started, the actual number of patients in this subgroup was 148 and this number is compared to the recovered number from the 18 weighted patients. At the start of the follow-up, this recovered number was 265, clearly overestimating the actual number. The figure shows how these two numbers (the actual and the recovered) evolve during the entire follow-up time. With the exception for the first year, the recovered and actual numbers are quite close to each other.

7.1.1.3 Question 3

> What is the impact of the choice of data which are reused to analyse the new outcome (when a choice is possible) on the final data set and for the interpretation of the estimates? (Study I)

The answer to question 3 is based on the number of patients which could be included in our final analysis (see Figure 4.1). The consequence of the alignment procedure in Study I was that the data used for analysis was reduced drastically: from 853 contralateral breast cancer cases, only 106 remained eligible, mainly due to the restriction of the study period. The maximum follow-up time period of the 'Metastases study' was nine years while the contralateral breast cancer cases were followed up during 30 years from breast cancer diagnosis date. This significant difference in follow-up length was the main reason why the cases data set was reduced so dramatically. However, as the criteria for both studies had further differences, the number of patients retained in the analysis was further reduced. As another consequence, the interpretation of the analyses results is limited to the population represented by the study base relevant to the current study, i.e. restricted to the same common criteria.

The main lesson which results from this observation is that, should there be the possibility to choose among several candidates for data sets to be reused, the best candidate will be characterised with larger criteria, if no other feature influences the choice.

7.1.2 Breaking the matching and weighting individuals

In general terms, breaking the matching and reusing data are similar. As in all studies, the matching was ignored, several questions were addressed in different studies and can be answered in a more general context than the context of the former section.

7.1.2.1 Question 4

> How well does the weighting system reconstruct the cohort and each subgroup of the cohort? (Studies III and IV)

Question 4 was addressed in detail in Study IV, and illustrated in Study III. We showed in Study IV that the Kaplan-Meir weighting system was able to reconstruct, on average, the correct number of individuals at risk over time. This was shown for the whole cohort, but also for all subgroups of the cohort. In cases of stratified sampling, the same results were obtained with stratified weights, while unstratified weights were not able to reconstruct the cohort appropriately. This is shown in Figure 7.3 obtained in simulation studies. Figure 7.4 shows how this did apply in Study III.

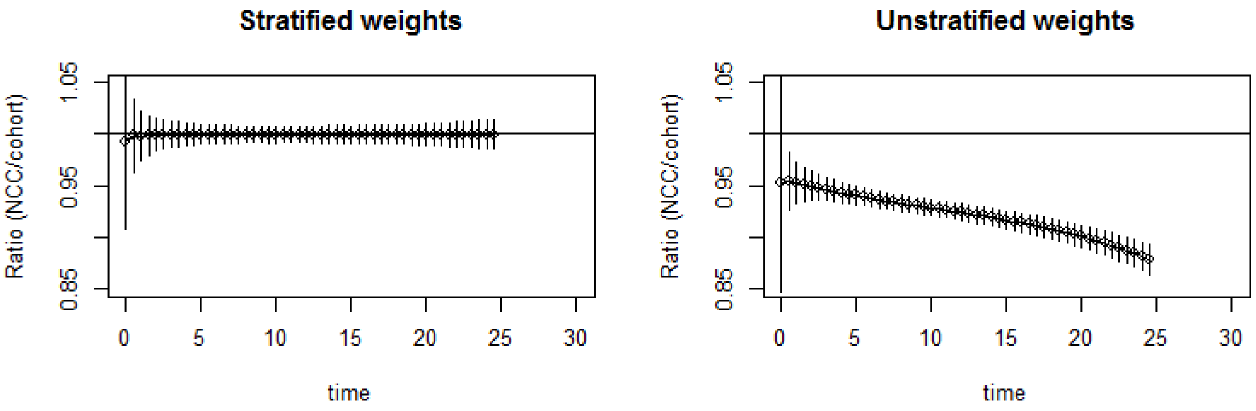

Figure 7.3: Ratio (with two standard deviations) of the numbers of individuals at risk over time recovered with stratified weights (left panel) and unstratified weights (right panel) from nested case-control data that was matched on the confounder, to the actual numbers of individuals at risk in the cohort

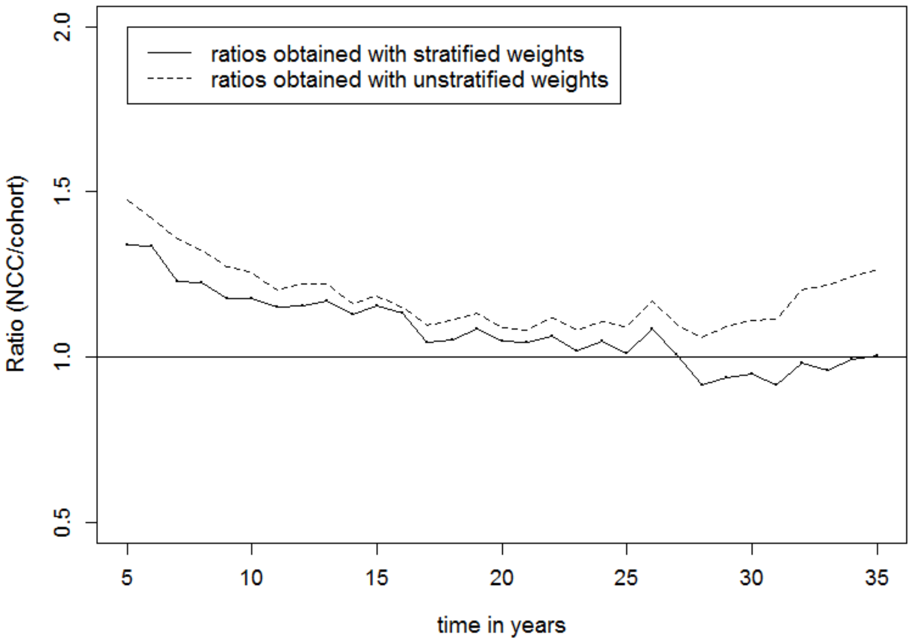

Figure 7.4: Ratio of the numbers of individuals at risk over time recovered with stratified weights in Study III, to the actual numbers of individuals at risk in the study base

In Figure 7.4, the recovered numbers over-estimated the actual numbers of individuals during the first third of the follow-up time. Several hypotheses can help to understand the reason for this. 1) Caliper matching was performed on age with an interval around the case's age of 5 years, while to compute the weights ten-year fixed age categories were used. The choice of ten-year categories broadened the risk sets to sample in, lowering the sampling probabilities, and hence increasing the weights. 2) The sampling, which was stratified on decade of breast cancer diagnosis dates could have been performed in a somewhat more convenient way than was planned, and could have been tighter that described. For example, if patients with diagnosis dates closest to the case's diagnosis date were more likely to be sampled; if this happened, it would have the same result as for the age variable above.

More important though, is the question regarding the impact of a poor reconstruction of the cohort on the estimates in the analysis. As the variables which could have led to the inaccurate reconstruction (age and decade of breast cancer diagnosis) were included in the analysis, and in the weights, we do not think that a bias is to be expected. This will further be illustrated in addressing the question which follows.

7.1.2.2 Question 5

> What is the impact of using unstratified weights instead of stratified weights to estimate covariates coefficients in matched designs? (Studies I, III and IV)

In Studies I, III and IV, the findings of Støer et al. [97] were challenged, according to which, stratified and unstratified weights provide similar estimates for the exposure of interest. In the fifth column of Table 7.1 the results from the analysis with unstratified weights are presented. The estimates for the exposures are similar to the estimates with stratified weights, provided that the potential confounders which were matching variables were included in the weighted Cox model. The main advantage of using stratified weights, is that we expect recovering correct estimates for all variables, including the matching factors used in the 'Metastases study' (Table 7.1, column 2), while using unstratified weights led to incorrect coefficient estimates for these latter coefficients (Table 7.1, column 5). This was further illustrated in Study IV, where we showed that unstratified weights yielded a correct estimate of the main exposure but were not correctly estimating the matching factor, whereas, with stratified weights, coefficients for both exposure and matching factors were correctly estimated (Figure 7.5).

The accuracy of exposure estimates with unstratified weights was also apparent in the results provided in Study III. They are presented in Table 7.2 and discussed later in the text.

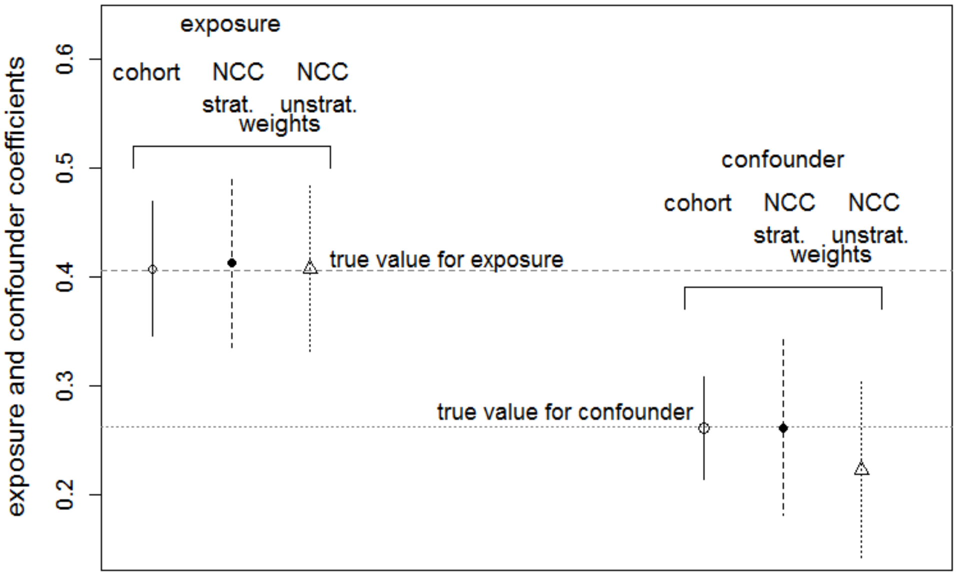

Figure 7.5: Estimates (± 1 standard error) for the exposure and the matching factor (confounder) obtained with the cohort analysed with Cox regression and the nested case control (NCC) data analysed with weighted Cox regression, using either stratified (strat.) or unstratified (unstrat.) weights.

7.1.2.3 Question 6

> What is the impact of using an appropriate method rather than a naïve statistical analysis when reusing nested case-control data? (Study I)

The sixth column of Table 7.1 provides an answer to this question. The estimates were biased when using a naïve unweighted analysis, compared to the weighted analysis in the second column of Table 7.1. It is difficult to judge in advance the extent of the bias for the main exposure(s), but serious bias in the matching factors could be expected. In this study, while the risk factor estimates were only slightly biased to the null, the estimates for age and adjuvant treatment (the matching factors in the prior study) were more seriously biased, as expected.

7.1.2.4 Question 7

> What is the impact of using selection dates of the controls when the censoring date in the nested case-control data set is ignored? (Study I)

Ignoring the last date of follow-up for the controls and using their selection date instead leads to important biases in the estimates, as the chosen date influences both the weight (through Equation 5.1 or 5.2) and the likelihood (through Equation 5.3) (data not shown). This highlights the care that has to be taken when collecting dates in nested case-control studies, should the collected data be intended for use in another study, or the IPW approach be used for the analysis.

7.1.2.5 Question 8

Is there a practical advantage to using Kaplan-Meier or glm/gam weights? (Studies I and III)

This question could be re-written as: Was it possible to choose the type of weights between Kaplan-Meier and glm/gam type of weights? Using the glm/gam type of weights was not possible with the anonymised data sets (Study I and III). Indeed, as the calculation of the glm/gam weights requires having a variable in the cohort data set indicating who was sampled for the nested case-control study, and as such indicators did not exist for ethical reasons, these weights could not be used. In order to highlight this statement, Figure 7.6 (inspired from Figure 4.1) illustrates the difference between the two possible situations. On the left hand side of the graph is a situation where the available data is not anonymised, which allows considering the available data as part of the study base, and having access to a broader choice for the type of weights, the statistical analyses and the software programs. On the right hand side is the situation of the data in Study I, where the data had been anonymised so that we can no longer link the available data to the study base and the choices are restricted, for weights calculation, statistical analyses and software programs.

A finding in this study was that the Kaplan-Meier weights are more flexible than the glm/gam weights as they can adapt to any situation. This advantage had never been mentioned before. Unfortunately, dealing with anonymised data hinders the (straightforward) use of the *multipleNCC* package developed in R, which means that the codes posted on http://www.meb.ki.se/~biostat/ are still useful.

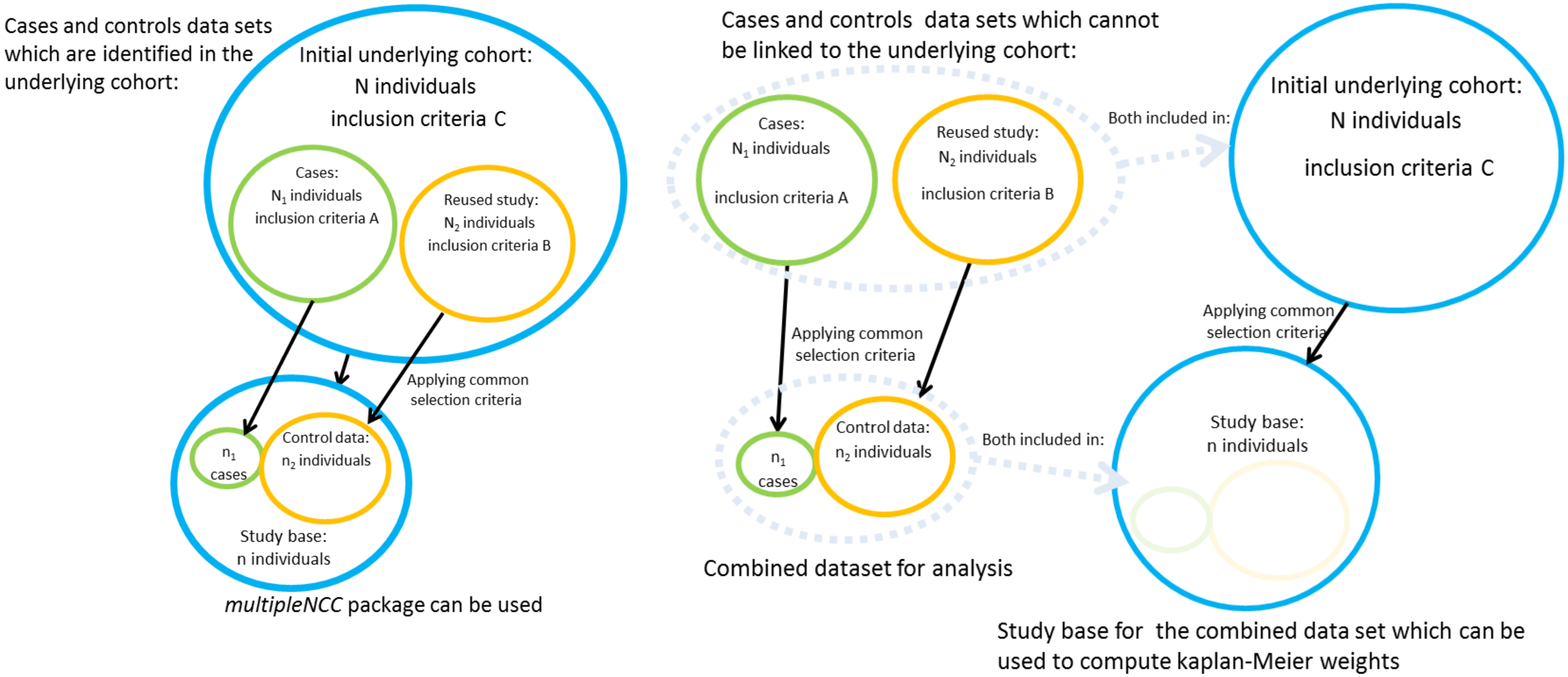
Cases and controls data sets which are identified in the underlying cohort:
Initial underlying cohort:
N individuals
inclusion criteria C
Cases:
N_1 individuals
inclusion criteria A
Reused study:
N_2 individuals
inclusion criteria B
Applying common selection criteria
n_1 cases
Control data:
n_2 individuals
Study base:
n individuals
multipleNCC package can be used
Cases and controls data sets which cannot be linked to the underlying cohort:
Cases:
N_1 individuals
inclusion criteria A
Reused study:
N_2 individuals
inclusion criteria B
Both included in:
Initial underlying cohort:
N individuals
inclusion criteria C
Applying common selection criteria
Applying common selection criteria
n_1 cases
Control data:
n_2 individuals
Both included in:
Study base:
n individuals
Combined dataset for analysis
Study base for the combined data set which can be used to compute kaplan-Meier weights

Figure 7.6: Difference between data which are included in the study base (with an identifier) (left panel) and data which do not have such identifier due to anonymisation (right panel). The figure is inspired from Figure 4.1 which applied to Study I.

7.1.3 Estimating absolute risk with case-control data

7.1.3.1 Questions 9 and 10

> Are the Langholz-Borgan and weighted methods suitable for estimating absolute risk from matched nested case-control data? (Study II)
>
> Are the Langholz-Borgan and weighted methods comparable in terms of accuracy and efficiency? (Study II)

These two questions were addressed in Study II and the results are shown in Figure 7.7. The Langholz-Borgan and weighted approaches both yielded unbiased absolute risk estimates at various time points (up to 20 years in the simulation studies) in the three following settings: (a) no additional matching, (b) matching on one confounder (gender), and (c) matching on two confounders (gender and age-group). The standard error of the Langholz-Borgan method was slightly larger than the standard error of the weighted method when matching was involved. We can thus confirm that the nested case-control design is an alternative to the cohort design for absolute risk estimation, and that it is possible to accommodate the Langholz-Borgan method in a matched nested case-control study. In addition, we found that there was no advantage of any of the two approaches in term of precision in case of no matching. In cases of matching, the weighted method was more precise than the Langholz-Borgan method.

In cases of fine matching (d), the weighted method should be preferred but, depending on how fine the matching is and how small the initial cohort is, there could be some bias with this latter method as well. In the simulation studies, the Langholz-Borgan approach gave biased absolute risk estimates at various time points and larger standard errors (especially for long-term prediction), while the weighted approach provided unbiased absolute risk estimates unless the initial cohort was small (10,000 individuals).

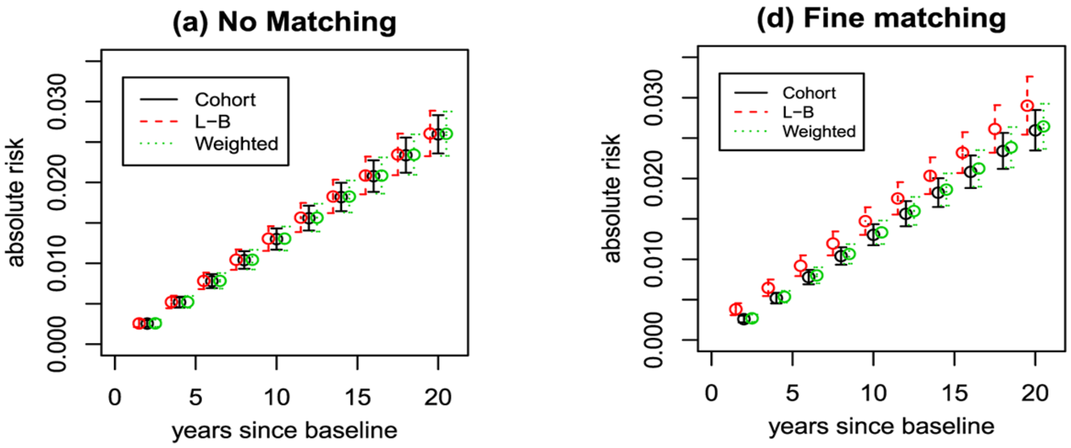

Figure 7.7: The average of absolute risk estimates for females across 500 realisations and their 95% confidence intervals in scenarios (a) no-matching and (d) fine matching. Cohort estimates (solid lines), Langholz–Borgan (L-B) estimates (dashed lines) and weighted estimates (dotted lines).

In the real data application, the weighted method performed better than the Langholz-Borgan method which provided biased estimates of absolute risk for long term prediction. The reason for this is not fully understood.

7.1.4 Overmatching

7.1.4.1 Question 11

Does weighted partial likelihood help overcome overmatching? (Study III)

This question was tackled in Study III. The estimates obtained in the analyses of the 1051 patients were similar in magnitude for both the conditional logistic regression and the weighted Cox regression, but the standard errors of the latter analysis were smaller (or equal) for all analyses (Table 7.2, two first columns). The overmatching was mitigated but the gain in power was quite modest by using weighted Cox regression instead of conditional logistic regression. As mentioned earlier, using stratified or unstratified weights led to similar estimates for the exposure coefficients (Table 7.2, columns 2 and 3).

Table 7.2: Adjusted coefficients (log hazard ratio) with standard errors for developing lung cancer five years or more after breast cancer. The conditional logistic regression was performed with 1018 patients in 509 matched sets and the weighted Cox regression is performed with 1051 unique patients.

Risk factors	Conditional logistic regression	Weighted Cox regression[a]	
		Stratified weights	Unstratified weights
	Log hazard ratio (standar error)		
	Univariate		
No radiotherapy	1	1	1
Radiotherapy	0.19 (0.16)	0.20 (0.16)	0.17 (0.15)
No smoking	1	1	1
Smoking	1.73 (0.19)	1.94 (0.16)	1.86 (0.16)
	Multivariable + interaction		
No radiotherapy and no smoking	1	1	1
Radiotherapy	-0.003 (0.29)	-0.26 (0.25)	-0.23 (0.24)
Smoking	1.37 (0.32)	1.38 (0.29)	1.37 (0.28)
interaction	0.54 (0.39)	0.78 (0.34)	0.68 (0.33)

[a] all weighted analyses included adjustment for the matching variables used in the sampling i.e. age (continuous), region and decade of diagnosis.

7.1.5 Clustered data

7.1.5.1 Question 12

> What are the advantages of using a weighted Cox regression with clustered data? (Study III)

In study III, using a weighted Cox regression, we were able to use all the information which had been gathered for the two lungs in each patient, thereby doubling the number of data observations and increasing the power of the statistical analysis. The approach used showed how to best exploit data which were collected on cases and controls, treated and non-treated with radiation therapy and with all doses reconstruction carried out for both lungs.

Some alternative to our approach exist. To use a conditional logistic regression analysis, while keeping the same outcome definition and retaining as much as data as possible, the data could have been been used as follows: the affected lung of a patient case is selected as case and analysed in sets with the ipsilateral lung of the control patient. This means that half of the data would be ignored, which would decrease the power as compared with our approach.

Another possibility is the approach used by Prochazka et al.[53] In their study, only women who were lung cancer cases treated with radiation therapy were selected, and the unaffected lung was considered as the control for the affected one. While this approach will avoid any problem of unmeasured confounding at the patient level and is definitely cost-efficient, it does not use the information from non-irradiated cases or control patients, and thus loses power.

7.1.6 Subgroup analyses

7.1.6.1 Questions 13 and 14

> Are subgroup analyses of nested case-control data valid when a subgroup is defined by a covariate measured at baseline? (Study IV)
> Is one method to be preferred among weighted Cox regression and conditional logistic regression for subgroup analyses of nested case-control data? (Study IV)

The results of Study IV provided clear answers to these two questions. As shown in Figure 7.8, both conditional logistic regression and weighted Cox regression provided on average an unbiased exposure estimate in subgroup analyses, disregarding the type of covariate used to define the subgroup (independent risk factor, matched or unmatched confounder, or effect modifier). The same conclusion applied in all tested simulation settings.

However, the conditional logistic regression presented such a variability compared to the weighted Cox regression, that the latter should be preferred. This is not surprising because, in weighted Cox regression, all individuals belonging to a subgroup are participating in the analysis, while, for the conditional logistic regression, only sets which, by chance, include a case and a control belonging to the same subgroup are used in the analysis. The number of sets which are lost will depend on the association between the outcome and the covariate and the prevalence of the various levels of the covariate as well as the prevalence of the exposure.

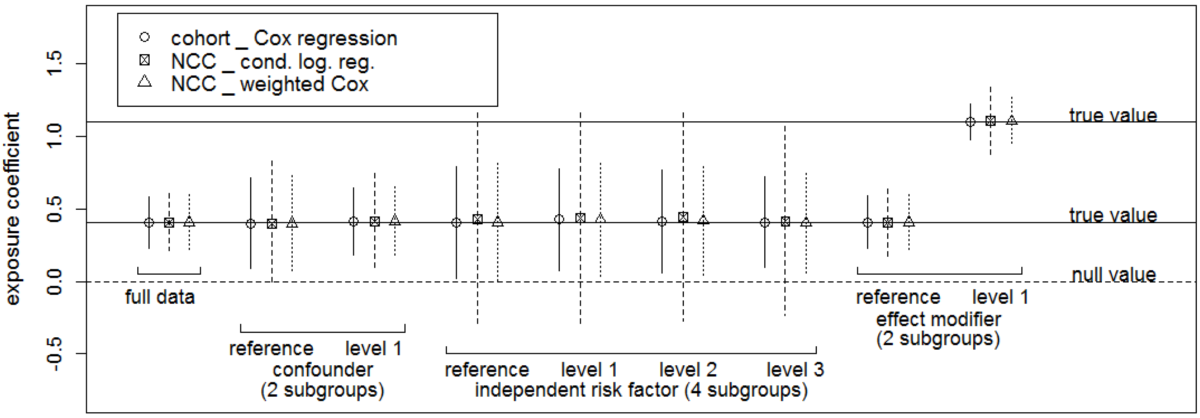

Figure 7.8: Adjusted exposure coefficient with 95% confidence intervals (calculated with model based standard errors), provided by the analysis of the cohort (Cox regression), the nested case-control data (conditional logistic regression and weighted Cox regression), the eight cohort subgroups (Cox regression) and the eight subgroups of nested case-control data (conditional logistic regression and weighted Cox regression).

7.2 Results regarding the epidemiological questions

7.2.1 Study I

The results regarding the epidemiological question of Study I (risk factors for contralateral breast cancer) are presented in the first column of Table 7.1. We found that the multifocality of the breast cancer tumour was a risk factor for contralateral breast cancer with a hazard ratio (95% confidence intervals) of 1.99 (1.07, 3.70). This is the first time that this factor is highlighted. This result is not surprising as multifocality has been shown to be associated with the lobular histological type of breast cancer, a known risk factor for contralateral breast cancer.[82,83]

Parity was found to be protective against contralateral breast cancer (with a hazard ratio (95% confidence intervals) of 0.40 (0.18, 0.89)), which is in line with the literature where parity is often reported as a protective factor but usually with a non-statistically significant hazard ratio.[73,80,81]

In addition, the main weighted Cox regression analysis confirmed that the known risk factors ‘family history’ and ‘non-ductal histological type’ were associated with an increased risk of developing contralateral breast cancer with hazard ratios (95% confidence intervals) of 1.91 (1.11, 3.28) and 2.09 (1.21, 3.59), respectively, in line with the literature.[71,73-77]

In the analysis, cases of contralateral breast cancer whose breast cancer was diagnosed between 1992 and 1997 were included in order to reduce the drastic loss of available case patients resulting from the study period alignment procedure (see Figure 4.1), or in other words to gain power. The same risk and protective factors were found when the time period for including contralateral breast cancer cases was 1992-2005 or 1997-2005, but the estimates were not statistically significant for the latter period. In our situation, there was no reason to suspect a problem by extending the study period to include contralateral breast cancer cases from 1992 (instead of 1997), as there were no major changes in this period for breast cancer diagnosis or treatment. This kind of choice is highly dependent on the clinical context and should be considered carefully.

7.2.2 Study III

In analysing how the risk of developing lung cancer after breast cancer is related to the dose of radiation therapy received for breast cancer treatment and the smoking habits of the patient, the following results were obtained.

An interaction was found between the two carcinogens (radiation therapy and smoking), meaning that the hazard ratio of developing lung cancer increased (doubled) in smokers receiving radiotherapy compared to smokers who were not treated with radiotherapy. If this had already been mentioned,[52,53,87] it is the first time that the interaction between smoking and breast cancer radiation therapy is characterised with a numerical value: the interaction coefficient was 0.78 (standard error = 0.34) leading to a hazard ratio of 2.19 (Table 7.2, second column).

The analysis of the 2102 lungs was performed with weighted Cox regression, which can handle the clustered data in a simple way by using a robust variance. Both hazard ratio and absolute risk for a lung to develop cancer were estimated. The results show that the risk of developing lung cancer among smokers increased with increasing radiotherapy dose (*P* for trend = 0.026), with a hazard ratio of 8.63 (95% confidence interval: 5.04, 14.79) for smokers who received a radiation dose higher than 13 Gy compared to a hazard ratio of 4.09 for smokers who did not have radiotherapy. In contrast, no such relationship was found among non-smokers.

The same observation applies when drawing the estimated curves of absolute cumulative risk for a lung cancer over the period from 5 to 25 years after breast cancer diagnosis (Figure 7.9)

The main findings from the study are the interaction between the two carcinogenic factors and the trend of increasing risk with increasing radiation dose in smokers. However, the clinical relevance is limited by features of the data: the individual dose reconstruction procedure was subject to inaccuracies,[53] the information on smoking was a simple binary variable,[115,116] and the confidence intervals in the analysis were rather wide. In addition, the risk of lung cancer was likely overestimated because death was treated as a censoring event and not as a competing outcome.

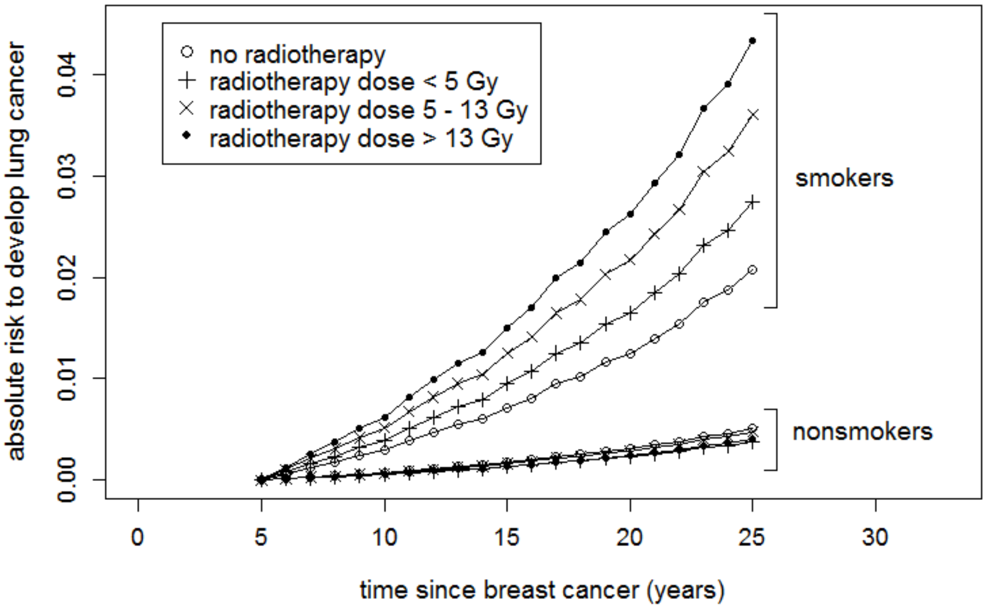

Figure 7.9: The estimated absolute risk (i.e., probability) of cancer in a lung exposed to various radiation doses, estimated at various time points from 5 to 25 years after breast cancer for patients aged 54 years at breast cancer diagnosis, assuming no competing risk of death.

8 CONCLUSIONS AND PERSPECTIVES

8.1 Achievements

Using both real data sets and simulation studies, we challenged the weighted partial likelihood approach in analysing nested case-control data, in order to better identify its advantages and limitations. In doing so, the scope of this approach was both refined and extended.

We showed that the weighting system using Kaplan-Meier weights enabled the reconstruction of the cohort, with, on average, the correct number of individuals at risk over time, both for the whole cohort, and in subgroups defined by a covariate measured at baseline. This result applied even for stratified design (i.e. when additional matching was performed on other covariates than time), provided that stratified weights were calculated. To our knowledge, this had never been shown before. This result suggests that any analysis or development made for cohort data could be applied to nested case-control data when using the IPW approach, but this would require confirmation studies.

We showed also how this result applied with some limitations to real situations, which highlights the importance of recovering the exact sampling scheme of the available data as well as the correct identification of the study base from which the nested case-control sampling was performed.

In case of stratified sampling, Støer et al. [97] already showed that unstratified or stratified weights provided similar exposure estimates if adjustments on the matching factors were made in the analysis. These results were confirmed in all the studies included in this thesis. In addition, when stratified weights were used, correct estimates for the matching factors were retrieved from the weighted partial likelihood, which is a further advantage of the IPW approach.

Regarding absolute risk estimation, we showed that both the Langholz-Borgan and the weighted methods provided valid estimates in most situations, the latter showing slightly higher levels of precision than the former. In case of fine matching, the Langholz-Borgan method was more prone to be biased than the weighted method and had larger standard errors, so that our recommendation would be to give preference to the weighted method for calculating absolute risk.

When overmatching is mentioned in the literature, it is often to try to prevent the issue from arising [54,57] and less frequently to explain how to mitigate the problem once it has arisen.[58] Using weighted partial likelihood to mitigate the problem of overmatching at the design stage is also part of the contribution of this thesis to the field, although results did only partially live up to the expectations.

The importance of using appropriate methods compared to a naïve unweighted approach which leads to biased estimates was also highlighted, which was already pointed in case-control studies.[33,34,54,117-119] We also demonstrated how the collected data on paired organs (i.e. clustered data) can be best exploited when using weighted partial likelihood.

The subgroups analyses shed light on the validity of such analyses and on the advantages of weighted partial likelihood compared to conditional logistic regression for the precision of the exposure estimate. It is reassuring that these analyses are valid per se, but we agree with Pocock [67] that subgroup analyses should not be overused, a topic addressed in the literature about randomised clinical trial.[65-69] Only subgroups defined by a covariate measured at baseline were investigated in our simulation studies, but extension to time-dependent covariates should not be an issue.[120]

We finally addressed practical aspects related to the reuse of data. The consequences of reusing data which have narrow inclusion criteria, the restriction in the choice of type of weights which can be calculated when data sets are anonymised, and the importance of having information on censoring dates for controls were all pointed out.

A major obstacle for using novel methods is the lack of software and guidelines. Study I provides such guidelines where all steps needed for data preparation and data analysis are explained in detail. Regarding software demand, an *R* package was developed and made available on CRAN,[110,111] but, due to the features of our available data sets in Studies I and III, the *multipleNCC* package could not be used without some adaptation. The codes which we provided (at http://www.meb.ki.se/~biostat/), include weights calculations which are performed with unique codes, but also with the *multipleNCC* package.

In this thesis, we focused on the IPW approach with Kaplan-Meier weights. The other likelihood methods, which were briefly exposed in Chapter 5 and which include a complete cohort likelihood approach,[18,94] or a multiple imputation approach,[95] would not be an option with the clinical data which was handled (Sudies I and III), as the nested case-control data set had been made independent from the study base. In such a situation the missing data and the available data cannot be linked which is a major obstacle for using these methods. The same obstacle prevented the use of glm/gam weights for the weighted partial likelihood approach as well as the straight utilisation of the *multipleNCC* package.

8.2 Nested case-control design and other cohort sampling strategies

In terms of study design, this thesis focused on the nested case-control design. The development of this design answered a need to optimise the efficiency of epidemiological research, but other designs answer the same need. Among the designs which sample within a well-defined cohort, the case-cohort design introduced in 1986 by Prentice [121] is another example of successful development. In the case-cohort design, the sampling is performed at baseline (inclusive sampling), where a subcohort is randomly sampled from the study base. This design and the nested case-control design (incidence density sampling) share aspects not only in terms of sampling,[5-7] but also in terms of analysis. A similar form of weighted partial likelihood is used for case-cohort data analysis, with the advantage of weights which are easier to calculate compared to the weights used in the weighted partial likelihood developed for the nested case-control design.[2,17,122-126] Vandenbroucke and Pearce [6] presented a third way to sample within a cohort, i.e. exclusive sampling. This "extreme" case-control design has even been extended to 'more extreme' case-control design by Salim et al. [127] who developed a weighted partial likelihood approach for analysing the data, with weights which also use basic information from the underlying cohort.

As the common feature of these three designs is the existence of a well-defined cohort, and as the weights used in the respective weighted partial likelihoods include basic information from the cohort, it gives the opportunity to highlight the value of having access to such cohorts in research.

8.3 The value of population and health registers

In Sweden, as well as in other countries, national/regional health and population registers are used for research purposes: cases for a specific outcome are selected and controls are subsequently sampled following a sampling procedure chosen by the research team and granted by the Ethical Review Board. The collection of detailed and often expensive data for the sample is then performed.

The value of the registers for this selection and sampling role is well recognised, but there is an added value which is less known and which was highlighted in this thesis: the ability to use basic information from the register to perform non-traditional analyses. Indeed, the weighted method used in the thesis was relevant because we had access to basic information available in the well-defined underlying cohort and the applied sampling procedure was known. Lacking one of these aforementioned factors can hamper the use of the IPW method in nested case-control data analysis.

On the other hand, while these factors are necessary, they are not sufficient. When analysing nested case-control data with conditional logistic regression, it is not necessary to collect any information on dates (event or censoring), as this information is useless for the analysis. The nested case-control data may thus lack information about dates, which could also hamper the use of the IPW approach. It is, therefore, equally crucial that the dates which are registered in the cohort are collected in the sample. While these should be easy to retrieve, this represents a change in some habits regarding data collection, and this could take some time before becoming the standard.

8.4 Perspectives

The achievements of this thesis are part of a bigger picture in epidemiology, which could be summarised with a key phrase: the need for flexibility. When data collection is expensive, there is a crucial need to optimise the cost-efficiency of research projects, which means both performing smart sampling and enabling the reuse of the expensive information which was gathered. This is valid for clinical data retrieved from clinical charts, for example, but it is even more crucial for genomics with the expensive and large associated molecular data.

Flexibility already characterises the development of various non-standard designs or sampling strategies which occasionally appear in the epidemiological literature. Their development is generally driven by a particular need, or a particular data situation. Such examples of needs and associated designs include: savings in time, cost or lowering computing burden (for example the "Exposure Enriched Case-Control" design [128]), the targeting of informative individuals (for example, the counter-matching sampling of controls,[129-131] and the end-point design [132]), and designs and analyses which do not need access to the underlying cohort.[133,134] This list is non-exhaustive.

In this thesis, flexibility was present in combining data sets and in using weighted partial likelihood, as this method of analysis is much more flexible than the conditional logistic regression analysis. A possible extension of this work could be to explore how to extract more information from a prior nested case-control study. After the end of the original study, some of the controls will usually develop the disease. If information was available on when these controls developed the disease, it would be valuable to find out how to use this new information in order to update the estimates obtained in the prior study.

One possible area of further development lies in exploring how to gain flexibility by combining designs into 'hybrid' designs. This kind of idea has already been explored with case-control studies.[135,136] Given the similarities in the weighted likelihoods used in both nested case-control and case-cohort data analyses, possible hybrid designs could be created from these two designs. For example, when a subcohort is set up in order to address several research questions, it will likely be unnecessarily large to address a specific question. In this case, a nested case-control sampling within the case-cohort data could be superimposed, capturing a substantial part of the information available in the case-cohort data with a much reduced sample size. Another example could be, on the other hand, to augment case-cohort data with a nested case-control sampling when the subcohort does not include enough individuals (a situation which could happen when the subcohort has been followed up for a long time).

Flexibility will always be needed when data have been collected following a given sampling procedure and are planned to be reused to address other research questions. We started a collaboration with a research team who had collected data following an extreme case-control design in order to answer an initial research question.[137] They are now willing to address another research question using the same expensive data, but using a subgroup of the patients of the first study. In addition to this restriction, the new question involves a different starting time in the underlying cohort, and a different outcome and censoring definitions. Being able to exploit any information from the underlying cohort, as well as taking advantage of the similarities of the designs and of the variety of the potential statistical approaches is the key to overcoming design issues in such complex situations.

8.5 Concluding remarks

This thesis has contributed to encourage researchers to use non-traditional approaches in analysing data. The wide applicability and potential advantages of the weighted partial likelihood approach for nested case-control data analysis has been further documented. We hope that this will eventually become standard practice.

The thesis also contributed to building bridges between epidemiological research and statistical methods, as well as highlighting the close links between epidemiological study designs, which are too often artificially distinguished.

9 ACKNOWLEDGEMENTS

When our family moved to Sweden several years ago, I did not have any professional project. This start in a new life gave me the opportunity to make an old dream come true: pursuing a PhD, studying a subject in depth and broaden my knowledge in statistics. At the end of this four-year journey, it is my pleasure to acknowledge all those who participated in the adventure.

I would like first to thank my main supervisor, **Marie Reilly**. When I contacted her five years ago, she kindly answered my e-mail and we met in person a few days later. I recall Marie telling me: "Pursuing PhD studies enables testing the student's will to overcome obstacles". Well, the good thing, is that all obstacles that I may encounter in my professional life, will now seem so easy to overcome. Thank you, Marie, for giving me the opportunity to explore a research topic which stimulated so well my interest, for your enthusiasm and our fruitful scientific discussions.

I wish to acknowledge my co-supervisor **Kamila Czene** for her constant support and her availability for scientific discussion. It was great to know that your door was always open, including for crisis management.

I never met my co-supervisor **Agus Salim** in person, and this is something that I regret. Yet I learned a lot through our e-mails exchange and I owe him much of the progress I made in programming in *R*. Thank you for your effective guidance despite the distance.

My mentor **Lucie Laflamme** was always present when I asked her advice or just a chat. Your benevolence has been so comforting during these years. Thank you so much, Lucie.

Henrik Larsson, **Lisen Arnheim-Dahlström** and **Amelie Plymoth** for their role as study director and support to the PhD students. I cannot overemphasise how important your role is in the PhD process. **Paul Lichtenstein**, for his caring leadership as head of department and for fostering a good collegial atmosphere.

A special thanks to **Anna Johansson**, **Maria Schelin** (född Sandberg) and **Maya Alsheh Ali** for their so helpful critical comments and suggestions on the draft version of my kappa. Anna, you impressed and inspired me with your professionalism in every area you are involved in. Our discussions, the ideas we have shared are bright moments that I will remember gladly, as well as your empathy which was so valued. Maria, I really loved your supportive presence and all precious moments we spent together, talking about science and life. It was also a pleasure to have co-authored a paper with you. Maya, I appreciated a lot your clear way of thinking, your kindness and your humour.

My fellow students **Andreas Karlsson**, **Tobba Palsdottir**, **Elisabeth Dahlqwist**, **Suo Chen** and **Myeongjee Lee** with whom I shared room 5201. A special thanks to Andreas with whom I stayed the longest and who always remind me that we did a PhD for fun. Thank you all for the good moments we spent together.

Special thanks to **Flaminia Chiesa** who offered me lots of encouragement and advice. Your friendship is so precious to me.

To **Nathalie Støer**, **Isabelle Le Ray**, **Rose Bosire**, **Sarah Walid Alsaadi**, and **Chen Wang**. We have all been members of Marie's team and had pleasure to share good moments and discussions together. A special thanks to Nathalie whose scientific articles are among my favourite readings. I am very honoured to co-author a paper with you and hope it will be published soon.

To my co-authors among whom **Edoardo Colzani**, **Niels Hagenbuch**, **Giovanna Gagliardi**, **Per Hall**, **Michaela Prochazka**, **Michal Abrahamowicz** and **Linda Lindström**, for the valuable collaboration and interactions we had during the editing of the manuscripts.

In All Likelihood, I will never forget **Johan Zetterqvist**, **Xingrong Liu** and **Hannah Bower** for their great comradeship while deciphering the subtleties of the famous statistical modelling course.

To **Caroline Weibull**, **Frida Lundberg** and **Peter Ström**, who kindly accepted to serve on the examination board for the pre-dissertation seminar.

To **Marie Jansson** and **Camilla Ahlqvist**, for being so helpful for clarifying administrative information and answering all questions so kindly. To **Gunilla Sonnebring** who is always available and smiling. To **Frida Palmér Thisell** and **Frank Pettersson** who discreetly take such good care of all members of the Department. To **Vivekananda Lanka** who is always smiling and cooperative and the whole **IT** team for their availability to help solving any popping up problem with technical tools.

To the whole Biostat group, thank you for all the interesting seminars which participated in broadening my statistical knowledge. And to all current and former PhD students at our department, thank you for the successful working groups, and efforts to make the PhD student life easy and enriching.

A special thanks to the 2014 Christmas party team with **Erika Nordenhagen**, **Lennart Martinsson**, **Jessica Pege**, **Barbro Sandin**, **Björn Gidlund**, **Gunilla Nilsson Roos**, **Michael Broms**, **Jie Song**, **Bojing Liu**, **Tong Gong** and all others who also participated. We had so much fun. Thanks a lot.

I also acknowledge CancerFonden, CAN-AIM, Radiumhemmets forskningsfonder and Förening för Medicinsk Statistik, which financially supported my work and scientific presentations at conferences.

I am grateful to my friends who have been so cheerful in the good and more difficult moments, **Elin and Patrik**, **Camille and Fabian**, **Wendy and Gunnar**, **Yvonne and Björn**. Thank you for your company which is so precious and invigorating.

I will never have enough words to thank my sambo, **Marc Struelens** for his steadfast support in dealing with the big and small events of the past years. I love you.

And finally to my best fan, **Elsa Struelens**, who has been so gracious in accepting my absence when I started my thesis, and drew such a beautiful cover picture to symbolize its results. Elsa, you are my treasure.

10 REFERENCES

1. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Third Edition. Philadelphia USA: Lippincott Williams & Wilkins; 2008. 87-99.

2. Borgan O, Samuelsen SO. Nested Case-control and Case-Cohort studies. In:Klein JP, van Houwelingen HC, Ibrahim JG et al. Handbook of Survival Analysis. Boca Raton, USA: Chapman & Hall/CRC; 2013. 343-367.

3. Collett D. Modelling survival data in medical research. Boca Raton, USA: Chapman & Hall/CRC; 2003. 11-13.

4. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat.* 1982;10(4):1100-1120.

5. Vandenbroucke JP, Pearce N. Incidence rates in dynamic populations. *Int J Epidemiol.* 2012;41: 1472–1479.

6. Vandenbroucke JP, Pearce N. Case-control studies: basic concepts. *Int J Epidemiol.* 2012;41:1480-1489.

7. Knol MJ, Vandenbroucke JP, Scott P et al. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *Am J Epidemiol.* 2008;168:1073-1081.

8. Pearce N. Classification of epidemiological study designs. *Int J Epidemiol.* 2012;41:393-397

9. Miettinen O. Design options in epidemiologic research. An update. *Scand J Work Environ Health.* 1982;8 suppl 1:7-14

10. Rodrigues L, Kirkwood BR. Case-control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls. *Int J Epidemiol*. 1990;19(1):205-13.

11. Wacholder S. The case-control study as data missing by design: estimating risk differences. *Epidemiology*. 1996;7:144-150.

12. Pearce N. Analysis of matched case-control studies. *BMJ*. 2016;352:i969.

13. Hosmer DW and Lemeshow S. Applied logistic regression. New York. Wiley Series in probability and mathematical statistics, Wiley; 1989. 187-213.

14. Thomas DC, Addendum to: Liddell JR, McDonald JC, Thomas DC. Methods of cohort analysis: Appraisal by application to asbestos mining. *J R Stat Soc Series A.* 1977;140(4):469-491.

15. Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *Am J Epidemiol.* 1990;131(1):169-76.

16. Ernster VL. Nested Case-Control studies. *Prev. Med.* 1994;23:587-590

17. Borgan O, Samuelsen SO. A review of cohort sampling designs for Cox's regression model: potentials in epidemiology. *Nor Epidemiol.* 2003;3(2):239-248

18. Saarela O, Kulathinal S, Arjas E et al. Nested case-control data utilized for multiple outcomes: a likelihood approach and alternatives. *Stat Med.* 2008;27(28):5991-6008.

19. Mazlan-Kepli W, MacIsaac RL, Walters M, Bath PM, Dawson J; VISTA Collaborators. Interruption to antiplatelet therapy early after acute ischaemic stroke: A nested case-control study. *Br J Clin Pharmacol* 2017 Mar 16. doi: 10.1111/bcp.13290. [Epub ahead of print]

20. Khanafer N, Vanhems P, Barbut F, Luxemburger C; CDI01 Study group. Factors associated with Clostridium difficile infection: A nested case-control study in a three year prospective cohort. *Anaerobe*. 2017 Mar 6;44:117-123. doi: 10.1016/j.anaerobe.2017.03.003. [Epub ahead of print]

21. Kim YH, Her AY, Rha SW, Choi BG, Shim M, Choi SY, et al. Routine angiographic follow-up versus clinical follow-up in patients with multivessel coronary artery diseases following percutaneous coronary intervention with drug-eluting stents: a nested case-control study within a Korean population. *Coron Artery Dis* 2017 Mar 7. doi: 10.1097/MCA.0000000000000479. [Epub ahead of print]

22. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KGM. Advantages of the nested case-control design in diagnostic research. *BMC Med Res Methodol.* 2008;8:48.

23. Chen K. Generalized case-cohort sampling. *J R Stat Soc Series B.* 2001;63(4):791-809.

24. Ohneberg K,Wolkewitz M, Beyersmann J, Palomar-Martinez M, Olaechea-Astigarraga P, Alvarez-Lerma F, Schumacher M. Analysis of Clinical Cohort Data Using Nested Case-control and Case-cohort Sampling Designs. A Powerful and Economical Tool. *Methods Inf Med.* 2015;54(6):505-514.

25. Wacholder S. Practical considerations in choosing between case-cohort and nested case-control designs. *Epidemiology.* 1991;2(2):155-158.

26. Swedish National Cancer Register. Available from: http://www.socialstyrelsen.se/register/halsodataregister [Last accessed: 19 April 2017]

27. UK Biobank. Available from: http://www.ukbiobank.ac.uk/about-biobank-uk/ [Last accessed: 19 April 2017].

28. Life Gene. Available from: http://lifegene.ki.se/ [Last accessed: 19 April 2017].

29. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to bisphosphonates and risk of cancer: a protocol for nested case–control studies using the QResearch primary care database. *BMJ Open.* 2012;2(1): e000548.

30. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to bisphosphonates and risk of gastrointestinal cancers: series of nested case-control studies with QResearch and CPRD data. *BMJ.* 2013;346:f114.

31. Frayling T.M., Timson NJ, Weedon MN, Zeggini E,Freathy RM, Lindgren CM et al. A Common Variant in the *FTO* Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science.* 2007;316:889–894.

32. Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL et al. Common variants in the *GDF5-BFZB* region are associated with variation in human height. *Nat Genet.* 2008;40:198–203.

33. Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol.* 2009;33(3):256-265.

34. Yung G, Lin X. Validity of using ad hoc methods to analyze secondary traits in case-control association studies. *Genet Epidemiol.* 2016;40(8):732-743.

35. Reilly M, Torrång A, Klint A. Re-use of case-control data for analysis of new outcome variables. *Stat Med.* 2005;24(24):4009-19.

36. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *JASA.* 1952;47(260):663–685.

37. Cochran WG. Sampling Techniques (3rd edition). New York: Wiley; 1977.

38. Salim A, Hultman C, Sparén P et al. Combining data from two nested case-control studies of overlapping cohorts to improve efficiency. *Biostatistics.* 2009;10(1):70-79.

39. Samuelsen S. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika.* 1997;84(2):379-394.

40. Wilson PWF, D'Agostino RB, Levy D et al. Prediction of coronary heart disease using risk factor categories. *Circulation.* 1998;97(18):1837-47.

41. D'Agostino Sr RB. Cardiovascular Risk Estimation in 2012: Lessons Learned and Applicability to the HIV Population. *J Infect Dis.* 2012;205 Suppl 3:S362-367

42. Cook NR, Paynter NP, Eaton CB, Manson JE, Martin LW, Robinson JG et al. Comparison of the Framingham and Reynolds Risk Scores for Global Cardiovascular Risk Prediction in the Multiethnic Women's Health Initiative. *Circulation.* 2012;125(14):1748-56, S1-11.

43. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ.* 2012;344:e3318.

44. Lin DY. On the Breslow estimator. *Lifetime Data Anal.* 2007;13:471-480.

45. Langholz B, Borgan Ø. Etsimation of absolute risk from nested case-control data. *Biometrics.* 1997;53(2):767-774.

46. Wolkewitz M, Cooper BS, Palomar-Martinez M, Olaechea-Astigarraga P, Alvarez-Lerma F, Schumacher M. Nested Case-Control Studies in Cohorts with Competing Events. *Epidemiology.* 2014;25(1):122-125.

47. Darabi H, Czene K, Zhao W, Liu J, Hall P, Humphreys K. Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement. *Breast Cancer Res.* 2012;14(1):R25

48. Klein AP, Lindström S, Mendelsohn JB, Steplowski E, Arslan AA, Bueno-de-Mesquita HB et al. An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population. *PLoS One*. 2013;8(9):e72311.

49. Karlsson K, Aly M, Clements M, Zheng L, Adolfsson J, Xu J et al. A Population-based Assessment of Germline *HOXB13* G84E Mutation and Prostate Cancer Risk. *Eur Urol.* 2014 ;65(1):169-176.

50. Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, Ingelsson E et al. Risk Prediction Measures for Case-Cohort and Nested Case-Control Designs: An Application to Cardiovascular Disease. *Am J Epidemiol.* 2012 ;175(7): 715–724.

51. Govindarajulu US, Lin H, Lunetta KL, D'Agostino RB. Frailty models: Applications to biomedical and genetic studies. *Stat Med.* 2011;30(22):2754–2764.

52. Grantzau T, Thomsen MS, Væth M, Overgaard J. Risk of second primary lung cancer in women after radiotherapy for breast cancer. *Radiother Oncol.* 2014;111:366e73.

53. Prochazka M, Hall P, Gagliardi G, Granath F, Nilsson BN, Shields PG, et al. Ionizing radiation and tobacco use increases the risk of a subsequent lung carcinoma in women with breast cancer: case-only design. *J Clin Oncol.* 2005;23:7467e74.

54. Breslow NE and Day NE. The analysis of case-control studies. *IARC.* 1980. Available from: http://www.iarc.fr/en/publications/pdfs-online/stat/sp32/SP32_vol1-0.pdf [Last accessed: 19 April 2017]

55. Hansson L, Khamis HJ. Matched samples logistic regression in case-control studies with missing values: when to break the matches. *Stat Methods Med Res.* 2008;17(6):595-607.

56. Pike MC, Hill AP, Smith PG. Bias and efficiency in logistic analyses of stratified case-control studies. *Int J Epidemiol.* 1980;9(1):89-95.

57. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. III. Design options. *Am J Epidemiol.*1992;135(9):1042-1050

58. Brookmeyer R, Liang KY, Linet M. Matched case-control designs and overmatched analyses. *Am J Epidemiol.* 1986;124(4):693-701

59. Marsh JL, Hutton JL, Binks K. Removal of radiation dose response effects: an example of over-matching. *BMJ.* 2002;325:327-330

60. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*. 2009;20:488e95.

61. Devore EE, Warner ET, Eliassen H, Brown SB, Beck AH, Hankinson S, Schernhammer E. Urinary melatonin in relation to postmenopausal breast cancer risk according to melatonin 1 receptor status. *Cancer Epidemiol Biomarkers Prev*. 2017;26(3):413-419.

62. Kim G, Jang SY, Han E, Lee YH, Park SY, Nam CM, Kang ES. Effect of statin on hepatocellular carcinoma in patients with type 2 diabetes: A nationwide nested case-control study. *Int J Cancer*. 2017; 140(4):798-806.

63. Boursi B, Mamtani R, Haynes K, Yang YX. Pernicious anemia and colorectal cancer risk - A nested case-control study. *Dig Liver Dis*. 2016;48(11):1386-1390.

64. Liu HC, Yang SY, Liao YT, Chen CC, Kuo CJ. Antipsychotic Medications and Risk of Acute Coronary Syndrome in Schizophrenia: A Nested Case-Control Study. *PLoS One*. 2016;11(9):e0163533.

65. Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. *Am Heart J*. 2000 Jun;139(6):952-61.

66. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064-1069.

67. Pocock SJ., Assmann SE., Enos LE. and Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med*. 2002;21(19):2917–2930.

68. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*. 2005;365(9454):176-186.

69. Sun X., Briel M., Busse JW. et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ*. 2012;344:e1553.

70. Oxman A. Subgroup analyses - The devil is in the interpretation. *BMJ*. 2012;344:e2022

71. Bernstein JL, Lapinski RH, Thakore SS, Doucette JT, Thompson WD. The descriptive epidemiology of second primary breast cancer. *Epidemiology*. 2003;14(5):552-558.

72. Sandberg ME, Hall P, Hartman M, Johansson AL, Eloranta S, Ploner A, Czene K. Estrogen receptor status in relation to risk of contralateral breast cancer–A population-based cohort study. *PLoS One*. 2012;7(10):e46535.

73. Vaittinen P, Hemminski K. Risk factors and age-incidence relationship for contralateral breast cancer. *Int J Cancer*. 2000;88(6):998-1002.

74. Hemminki K, Ji J, Försti A. Risks for Familial and Contralateral Breast Cancer Interact Multiplicatively and Cause a High Risk. *Cancer Res*. 2007;67:868-870.

75. Vichapat V, Gillett C, Fentiman IS, Tutt A, Holmberg L, Lüchtenborg M. Risk factors for metachronous contralateral breast cancer suggest two aetiological pathways. *Eur J Cancer*. 2011;47(13):1919-1927.

76. Reiner AS, John EM, Brooks J, Lynch CF, Bernstein L, Mellemkjær L et al. Risk of asynchronous contralateral breast cancer in noncarriers of BRCA1 and BRCA2 mutations with a family history of breast cancer: a report from the Women's Environmental Cancer and Radiation Epidemiology Study. *J Clin Oncol.* 2013;31(4):433-439.

77. Vichapat V, Garmo H, Holmqvist M, Liljegren G, Wärnberg F, Lambe M et al. Tumour stage affects risk and prognosis of contralateral breast cancer: results from a large Swedish-population-based study. *J Clin Oncol.* 2012;30(28):3478-3485.

78. Schaapveld M, Visser O, Louwman WJ, Willemse PH, de Vries EG, van der Graaf WT et al. The impact of adjuvant therapy on contralateral breast cancer risk and the prognostic significance of contralateral breast cancer: a population based study in the Netherlands. *Breast Cancer Res Treat.* 2008;110(1):189–197.

79. Hartman M, Czene K, Reilly M, Bergh J, Lagiou P, Trichopoulos D et al. Genetic implications of bilateral breast cancer: a population-based cohort study. *Lancet Oncol.* 2005;6(6):377-382.

80. Reeves GK, Pirie K, Green J, Bull D, Beral V; Million Women Study Collaborators. Reproductive factors and specific histological types of breast cancer: prospective study and meta-analysis. *Br J Cancer*. 2009;100(3):538–544.

81. Poynter JN, Langholz B, Largent J, Mellemkjaer L, Bernstein L, Malone KE et al. Reproductive factors and risk of contralateral breast cancer by BRCA1 and BRCA2 mutation status: results from the WECARE study. *Cancer Causes Control.* 2010;21(6):839-846.

82. Tot T. Clinical relevance of the distribution of the lesions in 500 consecutive breast cancer cases documented in large-format histologic sections. *Cancer*. 2007;110(11):2551-2560.

83. Dedes KJ, Fink D. Clinical presentation and surgical management of invasive lobular carcinoma of the breast. *Breast Dis.* 2008;30:31-37.

84. Franklin J, Paus MD, Pluetschow A, Specht L. Chemotherapy, radiotherapy and combined modality for Hodgkin's disease, with emphasis on second cancer risk. *Cochrane Database Syst Rev.* 2005, Issue 4. Art. No.: CD003187.

85. Grantzau T, Overgaard J. Risk of second non-breast cancer after radiotherapy for breast cancer: a systematic review and meta-analysis of 762,468 patients. *Radiother Oncol.* 2015;114:56e65.

86. Berrington de Gonzalez B, Gilbert E, Curtis R, Inskip P, Kleinerman R, Morton L, et al. Second solid cancers after radiotherapy: a systematic review of the

epidemiological studies of the radiation dose-response relationship. *Int J Radiat Oncol Biol Phys* 2013;86(2):224e33.

87. Kaufman EL, Jacobson JS, Hershman DL, Desai M, Neugut AI. Effect of breast cancer radiotherapy and cigarette smoking on risk of second primary lung cancer. *J Clin Oncol.* 2008;26:392e8.

88. Mattsson B, Wallgren A. Completeness of the Swedish Cancer Register. Non-notified cancer cases recorded on death certificates in 1978. *Acta Radiol Oncol.* 1984;23(5):305e13.

89. The National Academy of Sciences (NAS). Biological Effects of Ionizing Radiation (BEIR). Health risks from exposure low levels of ionizing radiation: BEIR VII Phase 2. 2006. Available from: http://www.nap.edu/catalog/11340.html [Last accessed: 19 April 2017]

90. Maddams J, Parkin DM, Darby SC. The cancer burden in the United Kingdom in 2007 due to radiotherapy. *Int J Cancer*. 2011;129:2885e93.

91. Hankin JH, Stram DO, Arakawa K, Park S, Low SH, Lee HP, Yu MC. Singapore Chinese Health Study: development, validation, and calibration of the quantitative food frequency questionnaire. *Nutr Cancer* 2001;39:187–195.

92. Lee M, Rebora P, Valsecchi MG, Czene K, Reilly M. A unified model for estimating and testing familial aggregation. *Stat Med.* 2013; 32(30):5353–5365

93. Kim S, De Gruttola V. Strategies for cohort sampling under the Cox proportional hazard model, application to an AIDS clinical trial. *Lifetime Data Anal*. 1999;5:149-172

94. Scheike T, Juul A. Maximum likelihood estimation for Cox's regression model under the nested case-control sampling. *Biostatistics*. 2004;5(2):193-206

95. Keogh RH, White IR. Using full-cohort data in nested case-control and case-cohort studies by multiple imputation. *Stat Med*. 2013;32(23):4021-4043

96. Støer N, Samuelsen S. Comparison of estimators in nested case-control studies with multiple outcomes. *Lifetime Data Anal*. 2012;18(3):261-283.

97. Støer N, Samuelsen SO. Inverse probability weighting in nested case-control studies with additional matching - a simulation study. *Stat Med* 2013:32(30):5328-5339.

98. Salim A, Delcoigne B, Villaflores K, Koh WP, Yuan JM, van Dam RM, M. Reilly. Comparisons of risk prediction methods using nested case-control data. *Stat Med.* 2017;36(3):455-465.

99. Samuelsen SO, Ånestad H, Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. *Scand J Statist.* 2007;34(1):103-119.

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

101. Kim RS. Analysis of Nested Case-Control Study Designs: Revisiting the Inverse Probability Weighting Method. *CSAM.* 2013;20(6):455–466

102. Kim RS. A new comparison of nested case–control and case–cohort designs and methods. *Eur J Epidemiol.* 2015;30:197–207.

103. Saarela O, Kulathinal S, Karvanen J. Secondary Analysis under Cohort Sampling Designs Using Conditional Likelihood. *J Prob Stat.* 2012;Volume 2012: Article ID 931416.

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

105. Borgan O, Keogh RH. Nested case-control studies. Should one break the matching? *Lifetime Data Anal.* 2015;21(4):517-41

106. Suissa S. The quasi-cohort approach in pharmacoepidemiology. Upgrading the nested case-control. *Epidemiology.* 2015;26(2):242-246.

107. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med.* 2006;25(24):4279–4292.

108. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med.* 2005;24(11):1713–1723

109. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med.* 2013,32(23):4118–4134

110. Støer N, Samuelsen SO. multipleNCC package. CRAN. 2015. Available from: https://cran.r-project.org/web/packages/multipleNCC/multipleNCC.pdf. [Last accessed: 19 April 2017]

111. Støer NC and Samuelsen SO. multipleNCC: Inverse Probability Weighting of Nested Case-Control Data. *R J.* 2016;8(2):5-18.

112. Støer N, Meyer HE, Samuelsen S. Reuse of controls in nested case-control studies. *Epidemiology.* 2014;25(2):315-317.

113. Zhou QM, Zheng Y, Cai T. Assessment of biomarkers for risk prediction with nested case-control studies. *Clinical Trials.* 2013; 10:677–679.

114. Delcoigne B, Hagenbuch N, Schelin ME, Salim A, Lindström LS, Bergh J et al. Feasibility of reusing time-matched controls in an overlapping cohort. *Stat Methods Med Res.* 2016 Sep 21. pii: 0962280216669744. [Epub ahead of print]

115. Prochazka M, Hall P, Granath F, Czene K. Validation of smoking history in cancer patients. *Acta Oncol.* 2008;47:1004e8.

116. Rachet B, Siemiatycki J, Abrahamowicz M, Leffondre K. A flexible modeling approach to estimating the component effects of smoking behavior on lung cancer. *J Clin Epidemiol.* 2004;57:1076e85.

117. Jiang Y, Scott AJ, Wild CJ. Secondary analysis of case-control data. *Stat Med.* 2006;25(8):1323-39.

118. Sun W, Joffe MM, Chen J, Brunelli SM. Design and analysis of multiple events case-control studies. *Biometrics*. 2010;66(4):1220-9.

119. Tapsoba Jde D, Kooperberg C, Reiner A, Wang CY, Dai JY. Robust estimation for secondary trait association in case-control genetic studies. *Am J Epidemiol*. 2014;179(10):1264-72.

120. Essebag V, Platt RW, Abrahamovicz M, Pilote L. Comparison of nested case-control and survival analysis methodologies foe analysis of time-dependent exposure. *BMC Med Res Methodol.* 2005; 5(1):5.

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

122. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann Stat.* 1988;16(1):64-81

123. Barlow WE1, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. *J Clin Epidemiol.* 1999;52(12):1165-72.

124. Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K, for the MORGAM Project. Case-cohort design in practice – experiences from the MORGAM Project. *Epidemiologic Perspectives & Innovations.* 2007; 4:15

125. Onland-Moret NC, van der A DL, van der Schouw YT, Buschers W, Elias SG, van Gils CH et al. Analysis of case-cohort data: a comparison of different methods. *J Clin Epidemiol*. 2007;60(4):350-355.

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

127. Salim A, Ma X, Fall K, Andrén O, Reilly M. Analysis of incidence and prognosis from 'extreme' case-control designs. *Stat Med.* 2014;33(30):5388-5398.

128. Huque MdH, Carroll RJ, Diao N, Christiani DC, Ryan LM. Exposure enriched case-control (EECC) design for the assessment of gene-environment interaction. *Genet Epidemiol.* 2016;40(7): 570–578.

129. Langholz B, Clayton D. Sampling Strategies in Nested Case-Control Studies. *Environ Health Perspect.* 1994;102(8):47-51

130. Cologne J, Langholz B. Selecting Controls for Assessing Interaction in Nested Case-control Studies. *J Epidemiol*. 2003;13:193-202

131. Cologne JB, Sharp GB, Neriishi K, Verkasalo PK, Land CE, Nakachi K. Improving the efficiency of nested case-control studies of interaction by selecting controls using counter matching on exposure. *Int J Epidemiol*. 2004;33(3):485-492.

132. Yao Y, Yu W, Chen K. End-point sampling. *Stat Sin.* 2017;27:415-435

133. Salim A, Ma X, Li J et al. A maximum likelihood method for secondary analysis of nested case-control data. *Stat Med.* 2014;33(11):1842-1852

134. Saarela O, Hanley JA. Case-base methods for studying vaccination safety. *Biometrics.* 2015;71(1):42-52

135. Stang A, Jöckel KH. Appending epidemiological studies to conventional case-control studies (hybride case-control studies). *Eur J Epidemiol.* 2004;19:527-532.

136. Haneuse SJPA, Wakefield JC. Hierarchical Models for Combining Ecological and Case–Control Data. *Biometrics.* 2007;63:128–136

137. Sboner A, Demichelis F, Calza S, Pawitan Y, Setlur SR, Hoshida Y et al. Molecular sampling of prostate cancer: A dilemma for predicting disease progression. *BMC Med Genomics*. 2010;3:8.