SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
0(0) 1–12

DOI: 10.1177/0962280216669744
smm.sagepub.com
SAGE

Feasibility of reusing time-matched controls in an overlapping cohort

Bénédicte Delcoigne,[1] Niels Hagenbuch,[1] Maria EC Schelin,[2] Agus Salim,[3] Linda S Lindström,[1,4] Jonas Bergh,[1] Kamila Czene[1] and Marie Reilly[1]

Abstract

The methods developed for secondary analysis of nested case-control data have been illustrated only in simplified settings in a common cohort and have not found their way into biostatistical practice. This paper demonstrates the feasibility of reusing prior nested case-control data in a realistic setting where a new outcome is available in an overlapping cohort where no new controls were gathered and where all data have been anonymised. Using basic information about the background cohort and sampling criteria, the new cases and prior data are "aligned" to identify the common underlying study base. With this study base, a Kaplan–Meier table of the prior outcome extracts the risk sets required to calculate the weights to assign to the controls to remove the sampling bias. A weighted Cox regression, implemented in standard statistical software, provides unbiased hazard ratios. Using the method to compare cases of contralateral breast cancer to available controls from a prior study of metastases, we identified a multifocal tumor as a risk factor that has not been reported previously. We examine the sensitivity of the method to an imperfect weighting scheme and discuss its merits and pitfalls to provide guidance for its use in medical research studies.

Keywords

Nested case-control, secondary analysis, weighted Cox regression, Kaplan–Meier type weights, GLM weights, cost-efficiency

1 Introduction

A case-control study is conducted in a well-defined source population (which can be a cohort) by sampling individuals according to their outcome status: usually all cases of the outcome of interest are selected and a defined number of controls (perhaps matched with the cases) are randomly sampled.[1] In the nested case-control (NCC) design, the controls are also matched on time, by sampling a defined number of controls at each time a case occurs from those individuals who are still at risk at that time (i.e. the "risk set").[1–3] This combines the significant savings in cost and time of classical case-control studies with the benefit of the additional information in the time aspect of a cohort design. The NCC design provides cost-efficient unbiased estimates of hazard ratios for measured risk factors under the proportional hazards model, provided the controls are randomly selected at the case event times within the strata defined by the matching variables and that there is no survivor bias for the unexposed cases.[2–4] This efficiency has led to widespread use of the NCC design in non-communicable and infectious disease epidemiology.[5,6] Other applications include the evaluation of screening and vaccination programs.[7,8] As a result, there is a large volume of NCC data available for potential secondary analyses, which are of increasing interest in medical applications where data collection and/or measurement requires significant investment. However, the controls from NCC studies cannot be readily reused to address a new research question due to the matching

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[2]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[3]Department of Mathematics and Statistics, La Trobe University, Victoria, Australia
[4]Department of Surgery, University of California, San Francisco, CA, USA

Corresponding author:
Bénédicte Delcoigne, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Box 281, SE-17177 Stockholm, Sweden.
Email: benedicte.delcoigne@ki.se

on the time of occurrence of the case's outcome.[9,10] To overcome this weakness of the NCC design, novel approaches to reusing these controls have been developed in the last decade including the likelihood method of Saarela et al. and the weighted likelihood methods of Salim et al.[11–13] In the weighted likelihood, each reused control is weighted by the inverse of his/her probability of being sampled to correct the bias induced due to the matching on time in the prior study. All cases get a weight of one, reflecting the 100% certainty of being in the NCC study. In contrast, controls get a larger weight, which make them representative of the number of similar eligible individuals. The hazard ratios for risk factors can be obtained from a weighted Cox regression analysis.[14] The variance estimator developed by Samuelsen for these weighted estimates is not easy to implement.[15] On the other hand, this estimator was shown to be in good agreement with an easy to implement robust variance estimator that does not account for the uncertainty in the weights and that might lead to more conservative inference in some circumstances.[14,16] Such weighted analyses have been shown not only to be unbiased,[17–19] but also to have better efficiency than conditional logistic regression.[18,19]

The performance of the weighted method in reusing NCC data has been studied and documented through simulation studies and illustrative data analyses of simplistic scenarios in the medical literature.[12–14,20,21] All these studies showed an increased efficiency of the NCC design when one uses additional controls that were previously sampled to study another outcome/endpoint in the same underlying cohort.[12,21] Støer and Samuelsen[14] compared different ways of calculating the weights and showed that the choice of the type of weights does not matter much. Salim et al.[13] explored the feasibility and efficiency of a potential study that gathered no controls at all, but relied only on previously sampled data. While all these published studies agree on the gain in efficiency of reusing NCC data, the methods have been largely ignored by medical researchers and have not found their way into routine use by biostatisticians and epidemiologists.[22] There is no published work of an application of the method to a setting where the data have features typical of real research studies, such as cases and available controls being sampled from different (but overlapping) cohorts and with different inclusion and sampling criteria. Salim et al. and Støer and Samuelsen worked with both simulated data and real data, and had the advantage of being able to compare the estimates obtained with several methods, either because they had access to the full cohort (with simulation studies) or actually had controls sampled for the second (i.e. new) outcome which could be used in a validation step.[12–14,20,21] In our application, we go one step beyond Salim's work by discussing the pros and cons of reusing NCC data in a real situation (i.e. data gathered to address a real research question) where the second study did not gather any controls, so that the only option is to reuse prior control data. We also clarify the conclusion of Støer and and Samuelsen concerning the choice of weights, highlighting how the feasibility of calculating several types of weights may depend on the data at hand. We provide a step-by-step description of when and how one can perform a valid analysis and we compare the reuse of the sampled controls alone versus the reuse of the controls and cases of the prior outcome as the reference group for the new outcome.

In this paper, we address the reuse of individuals from a prior NCC study of a different outcome, in the context of anonymised data sets that do not have any identifier enabling their linkage to the cohort, as would be typical in practice: the prior NCC data set, the cases of the new outcome being studied and the basic information for the cohort from which these two were sampled are available in three separate data sets. We apply the weighted method using standard software commands, describing in detail the steps involved in the data preparation, calculation of the weights, and the handling of matching variables, and illustrating how each step is executed in our own application.

2 Methods

2.1 Setting and design

2.1.1 Three data sets to handle

We consider a situation where a NCC study was previously conducted in order to investigate risk factors for a defined outcome. This NCC study (which we will refer to as the prior NCC study) sampled cases of the outcome of interest (i.e. the prior outcome) and controls from a well-defined cohort. For the sampled individuals, data on risk factors of interest were collected, for example from interviews, questionnaires, medical charts or biological specimens. The data were anonymised to ensure data protection for the individuals, so that the subjects selected into the prior NCC study could no longer be linked to the cohort. In addition to information on dates (of entry, both outcomes of interest, censoring events and last follow-up), both the cohort and the NCC data set include the matching factors used to perform the sampling.

A second research question now arises concerning another (i.e. new) health outcome in an overlapping cohort, for which investigators wish to study several potential risk factors that are available in the prior NCC data.

We assume that cases for this new outcome have been selected, and after collecting information for the variables of interest the data have been anonymised.

2.1.2 The contralateral breast cancer study

The cohort: All Swedish breast cancer diagnoses in the Stockholm-Gotland health-care region are reported to the Stockholm Breast Cancer Register since 1976. In Sweden, such registers are used by researchers to identify cases of interest and to subsequently sample control patients before assembling additional details, for example from patient medical records in the hospitals. In our illustrative example, the underlying cohort from which all patients in the data sets were selected consisted of the 32,153 non-metastatic breast cancer (BC) patients who were registered in the Stockholm Breast Cancer Register from 1976 to 2008. In addition to recording the date of the BC diagnosis and some other limited information (Table 1), patients in this register are followed up for metastases, contralateral breast cancer, other malignancies, and death.

The prior NCC study: The "Metastases study" is nested in the Stockholm Breast Cancer Register described above and was conducted to investigate genetic risk factors for metastases following an invasive BC. In this study, BC patients from the register were eligible if they met the following criteria (Table 1, upper panel): a primary non-metastatic breast cancer diagnosed during the period 1997–2005; age at diagnosis of BC at most 75 years; chemotherapy and/or hormonal therapy prescribed as adjuvant treatment for the BC. Cases of metastases were defined as patients who developed metastases at any time after the BC and before 2006. For each case, metastases-free controls were selected who matched the case on time since breast cancer diagnosis, age category (<45, 45–54, >54 years), intended adjuvant treatment (chemotherapy; hormonal therapy; a combination of the two), and a dichotomous variable indicating whether the BC was diagnosed before or after the end of 2000. In this study, 191 metastases cases were included for whom a total of 615 controls were sampled. Data were extracted from the medical records of these 806 patients for several covariates of interest, including parity, family history, histological type of the initial breast cancer, and BC tumor multifocality (Table 1, lower panel).

Table 1. Inclusion criteria and available variables for the initial cohort, the CBC cases and the patients in the Metastases study.

	Initial cohort	CBC cases	Metastases study
Sampling criteria			
First invasive BC diagnosis date	1976–2008	1976–2005	1997–2005
Age			$\leq$ 75 years
Intended treatment for first BC			Includes chemo- or hormonal therapy
CBC diagnosis date		$\geq$ 3 months after BC	
Other malignancies		No malignancy prior to BC	
Available variables			
Dates			
First BC diagnosis	Yes	Yes	Yes
Death	Yes	Yes	Yes
CBC diagnosis	Yes	Yes	Yes
Metastasis diagnosis	Yes	No[a]	Yes
Other malignancy diagnosis	Yes	No[a]	Yes[b]
Patient characteristics			
Age at first BC	Yes	Yes	Yes
Family history	No	Yes	Yes
Parity	No	Yes	Yes
Tumor characteristics and treatment of first BC			
Multifocality	No	Yes	Yes
Histological type	No	Yes	Yes
Intended adjuvant treatment	Yes	Yes[b]	Yes

[a] Not permitted by the inclusion criteria.

[b] Not available in the study data set but retrieved by merging with the initial cohort.

The new outcome: The new research question concerns risk factors for another (i.e. new) outcome, namely contralateral breast cancer. Contralateral breast cancer (CBC) cases are defined as women having non-metastatic invasive BC as a first malignancy, followed at least three months later by a non-metastatic invasive CBC as a second malignancy.[23] These cases are, by definition, included in the Stockholm Breast Cancer Register presented above. The original aim was to investigate risk factors for CBC using a NCC design. A total of 853 CBC cases were identified in the register and data on patient and tumor characteristics were extracted from the medical records, including tumor multifocality (which has not been previously studied) and three exposures (parity, family history and histological type of the initial breast cancer) that have been investigated in other published studies[24,25] (Table 1). Age and adjuvant treatment were considered as potential confounders.[26] When discussions about control selection were under way, it was realised that the required information on exposure and confounder variables was already available in the Metastases study. The main consideration was then the appropriate use of the 806 available study subjects as a reference (control) group for the CBC cases. The use of these data was approved by the Stockholm Regional Ethics Committee.

2.2 Preparation of the data sets

2.2.1 Alignment of the data sets

We propose to analyse cases of the new outcome by comparing them to a reference group of non-cases available from a prior NCC study, all of whom were selected from the same background population. We will refer to this reference group as the "available controls" to distinguish them from the "sampled controls" in the prior NCC study. The first step of the data preparation is the "alignment" of the data sets to identify the common/overlapping cohort, using the known inclusion/exclusion criteria. This overlapping cohort is the underlying study base from which the prior NCC data and the current cases were both sampled, so that the (current) cases and available controls can be combined in a valid analysis.

2.2.2 Alignment of the three data sets for the CBC study

Given that different criteria were used for selecting the CBC cases and sampling the patients in the Metastases study (Table 1, upper panel), we describe here how we aligned these data sets and identified a common study base from which the patients for both studies were sampled. To meet the inclusion criteria of the Metastases study, we restricted the cohort of BC patients to those whose BC was diagnosed between 1997 and 2005, were at most 75 years old at diagnosis and whose prescribed adjuvant treatment included chemotherapy and/or hormonal therapy (see Figure 1). To represent patients who could be selected for the CBC study, we further restricted to those with at least three months of follow-up and who had no prior malignancy. The resulting cohort is the underlying study base of 6867 patients from which all the cases and available controls for our analysis were sampled. Likewise, all CBC cases were restricted to patients who had their BC in the study period 1997–2005, were at most 75 years old at BC diagnosis and had received adjuvant treatment that included chemotherapy and/or hormonal therapy (Figure 2). In parallel, patients from the Metastases study (the available controls) were restricted to those without any prior malignancy (except BC) and with a three months minimum follow-up time from BC diagnosis (Figure 1).

2.3 Calculation of the weights

2.3.1 Kaplan–Meier type weights

To analyse the combined cases of the new outcome and the available controls from the prior NCC study using weighted Cox regression, each individual must be weighted by the inverse probability of being sampled for either study.[12] The sampling probabilities are computed using the overlapping study base, which must contain a minimum set of variables (i.e. entry and censoring/event dates for both outcomes) for all members. This basic information is referred to as the "skeleton" of the study base.[15] When there are matching variables (in the prior study), the skeleton must also contain this information. The sampling probability for an individual sampled as a control in the prior study depends on whether this individual is available to be sampled at one or more of the times of disease occurrence of the cases, the number of eligible controls at each of these event times (i.e. the size of the risk sets) and the number of controls who are sampled at each event time. The only one of these three components requiring computational effort is the size of the risk sets. As the main criterion to sample the controls in the prior NCC study was the elapsed time since a defined "time zero", the risk set sizes at each event time can be readily calculated from the study base using a Kaplan–Meier table. If, in addition to time, controls were matched

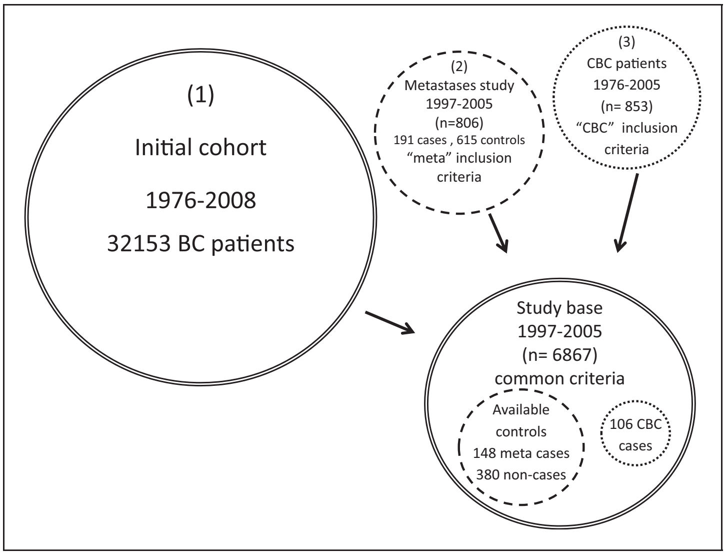

Figure 1. Alignment of the three data sets. The common criteria to align the three data sets were: study period 1997–2005, age ≤ 75 years old, intended adjuvant treatment including chemotherapy and/or hormonal therapy, no prior malignancy and at least three months of follow-up.

═══ (1) The initial cohort included 32,153 BC patients from 1976 to 2008. The alignment led to a study base of 6867 patients.

▬ ▬ ▪ (2) The initial data set of the Metastases study included 806 BC patients from 1997 to 2005 who were at most 75 years old, with prescribed adjuvant treatment including chemotherapy and/or hormonal therapy. After alignment, the data set contained 528 unique patients.

▪▪▪▪▪▪▪▪ (3) The initial data set of the CBC study included 853 CBC cases. Applying the common criteria led to a data set with 106 CBC cases.

on additional variables, a stratified Kaplan–Meier table provides the appropriate risk set sizes in each of the strata (see Appendix 1 for details).

The available controls get a weight which is the inverse of their sampling probability. By their weight, they represent all the similar individuals who could have been selected. This weight will be 1 for the prior cases who were sampled with certainty and > 1 for all the sampled controls except those few that were also cases, whose weight will be 1. Cases of the new outcome receive a weight of one to reflect the certainty of being selected. If a sampled control in the prior NCC study later became a case of the new outcome, they also enter the analysis as a case with weight 1.

2.3.2 Weight calculation in the CBC study

In our application, we have the skeleton of the study base (a register at county level) with the basic information on dates and matching variables that were used for the sampling in the Metastases study. For the CBC cases, the sampling probabilities and weights are equal to one since all cases were included. The cases in the Metastases study were also assigned a weight of one. The controls that were sampled for the Metastases study were selected to match metastases cases. To compute their probability of being sampled, the Kaplan–Meier table was run on our aligned study base using the metastases event times. To reflect the stratified sampling, we used a stratified Kaplan–Meier table with 18 strata defined by the levels of the three matching factors. The weights were then calculated as described in Appendix 1.

2.4 Statistical methods

2.4.1 Weighted Cox regression analysis

The main analysis of the potential risk factors for the new outcome uses a weighted Cox regression model with weights as described above. All patients are included from their time of entry into the common study base. Available controls are censored at the first of the following dates: censoring time (defined for the new outcome), death or end of the study. The hazard function for patient i is given by: $h_i\ (t|x_i,\, z_i,\, \boldsymbol{\beta},\, \boldsymbol{\gamma}) = h_0(t)\ \exp(\boldsymbol{\beta}\ x_i + \boldsymbol{\gamma}\ z_i)$

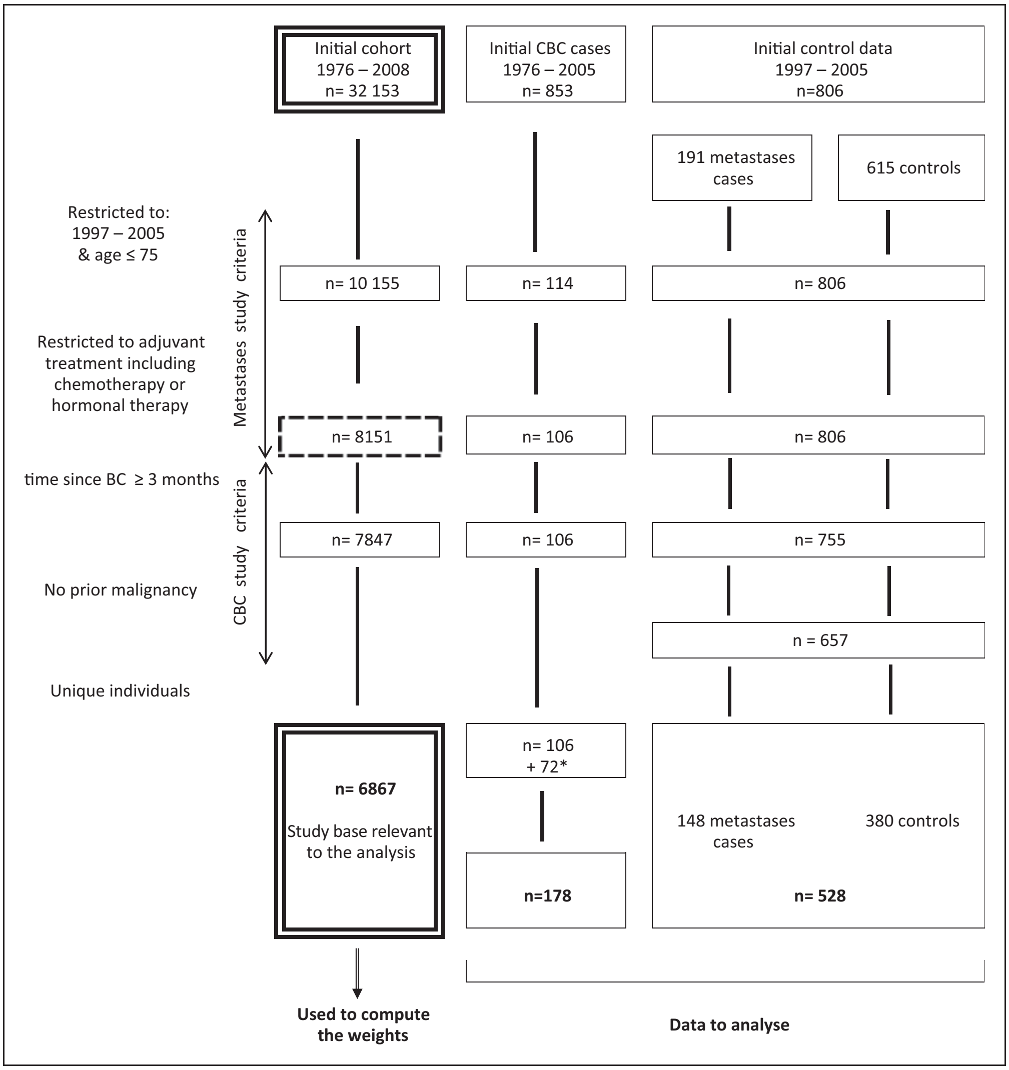

Figure 2. Flow chart showing how the different sampling criteria for the Metastases study and CBC cases influenced the resulting data sets.
*additional CBC cases in 1992-1996 included (see text).

where x_i is the vector of covariates of interest, z_i is the vector of potential confounders (which may include matching variables used in the prior study), and $\boldsymbol{\beta}$ and γ are the corresponding coefficients.

2.4.2 Weighted Cox regression analysis in the CBC study

In our application, the time of entry into the study base was three months after the initial BC diagnosis date, which is the "time 0" for all patients in the analysis. The available controls were censored at the first of the following dates: any malignancy (including metastasis), death or end of the study. Note that those available controls who developed metastases get a weight of 1 as they were sampled with certainty for the prior study. The exposures of interest in our study were tumor multifocality, parity, histological type and family history (i.e. up to third degree relative having BC), and the risks associated with these factors were to be adjusted for matching variables used in the prior study that were potential confounders (age and adjuvant treatment). We also conducted the weighted Cox regression analysis including only the sampled controls ($n = 398$) from the Metastases study. To measure the effect on the estimates due to an imperfect reconstruction of the study base, we re-ran the analysis with weights computed from the study base relevant only to the Metastases study, i.e. the study base which meets only the inclusion criteria of the Metastases study rather than the study base common to both the CBC and metastases patients. The results of our main analysis were also compared to those from a naïve (i.e. unweighted) Cox regression analysis.

2.5 Statistical software

All data management and analyses were performed with the R statistical package (version 3.0.2). To extract the Kaplan–Meier tables and run the Cox regression analyses, we used the *survfit* and *coxph* (with the option "*weights*") functions respectively, provided in the *survival* package. A robust variance estimator, which is readily available in standard statistical software, is used to estimate the standard errors of the estimates from the weighted analysis.

3 Results

3.1 Alignment procedure

3.1.1 General consequences of the alignment procedure

The alignment procedure described in section 2.2 restricts the three data sets to a common set of inclusion criteria. As a result, the underlying cohort, the prior NCC data set and the cases of the new outcome will usually contain fewer individuals.

3.1.2 Consequences of the alignment procedure for the CBC study

Restricting the underlying cohort to the study period 1997–2005 and to ages lower than 76 resulted in 10,155 patients, of whom 8151 fulfilled the criteria concerning treatment in the Metastases study (Figure 2, first column). Further restriction due to selection criteria used in the CBC study resulted in a final study base of 6867 patients. Fifty-one records from the control data set were excluded due to censoring within three months of their original BC diagnosis and a further 98 due to other malignancies prior to BC. Since the weighted Cox regression uses unique individuals,[13] any control who had a CBC during follow-up contributed as a CBC case and our control data set finally included 528 unique control patients (Figure 2, third column). The CBC cases were restricted with the same criteria, but with a five-year extension of the study period (from 1992), as there were no major changes in diagnosis nor treatment for breast cancer during this time. The final data available for analysis are from 706 unique patients, 178 CBC cases and 528 controls.

3.2 Weighted Cox regression analysis

3.2.1 General

The main analysis is a weighted Cox proportional hazards regression run on a data set in which the control data includes individuals (both cases and controls) from a prior NCC study. A naïve approach that conducts a Cox regression of these data without weighting will result in biased estimates. The accuracy of the estimates could depend on the ability to reconstruct the study base correctly. If data from a prior NCC study are available only for sampled controls, this will result in only minimal bias of the estimates, as the controls contribute most of the information to the analysis due to their larger weight.

3.2.2 Results for the CBC study

The estimates of our weighted Cox regression analysis are presented in the first column of Table 2 and suggest a protective effect of parity and an increased risk for women with multifocal tumors, non-ductal histological type or positive family history. Similar estimates were obtained when considering only the sampled controls from the prior NCC data set (Table 2, second column). The adjusted estimates of risk from our main analysis are similar to those obtained using weights computed from an imperfectly reconstructed study base (i.e. the study base relevant to the Metastases study) (Table 2, third column). The naïve unweighted analysis provided slightly biased estimates for the risk factors and strongly biased estimates for the confounders (Table 2, fourth column).

4 Discussion

This study presents in detail the application of weighted Cox regression in a situation where individuals from a previous NCC study serve as controls to address a new research question in an overlapping cohort. We have shown how to implement the method with simple commands available in standard statistical software in the realistic setting where data sets have been anonymised and do not have any common identifier with the cohort from which the individuals were sampled. Using the method to investigate risk factors for CBC, we obtained estimates consistent with the literature for a number of risk factors and found evidence of a higher risk of CBC for multifocal tumors, a risk factor that has not been previously reported.

Table 2. Adjusted risk estimates: hazard ratios (HR) and 95% confidence intervals (CI) from Cox regression analyses.

Risk factors	Main analysis[a]	Sampled controls[b]	Study base for metastasis[c]	Unweighted[d]
Non-multifocal tumor (ref.)	1	1	1	1
Multifocal tumor	1.99 (1.07, 3.70)	2.03 (1.08, 3.81)	1.98 (1.07, 3.70)	1.60 (1.06, 2.41)
Nulliparous (reference)	1	1	1	1
Parity	0.40 (0.18, 0.89)	0.37 (0.16, 0.85)	0.39 (0.18, 0.86)	0.79 (0.47, 1.33)
Ductal histological type (ref.)	1	1	1	1
Non-ductal histological type	2.09 (1.21, 3.59)	2.11 (1.22, 3.67)	2.11 (1.22, 3.66)	1.78 (1.24, 2.57)
No family history (reference)	1	1	1	1
Positive family history	1.91 (1.11, 3.28)	1.93 (1.12, 3.34)	1.92 (1.12, 3.32)	1.48 (1.04, 2.11)
Chemotherapy (reference)	1	1	1	1
Hormonal therapy	0.71 (0.39, 1.26)	0.72 (0.40, 1.31)	0.65 (0.36, 1.16)	1.94 (1.24, 3.05)
Chemo + Hormonal therapy	0.57 (0.30, 1.07)	0.57 (0.30, 1.09)	0.57 (0.30, 1.07)	0.73 (0.43, 1.26)
Age $<$ 45 (reference)	1	1	1	1
Age 45–54	1.23 (0.62, 2.44)	1.25 (0.62, 2.50)	1.16 (0.58, 2.32)	1.61 (0.96, 2.73)
Age $>$ 54	0.96 (0.49, 1.88)	0.94 (0.48, 1.85)	0.93 (0.48, 1.81)	1.17 (0.69, 1.96)

[a]Results obtained from a weighted Cox analysis, adjusted for assumed confounders and with weights computed by stratified Kaplan–Meier analysis.

[b]Same as main analysis with the 398 sampled controls only.

[c]Same as main analysis with weights computed with the imperfectly reconstructed study base, i.e. the study base relevant to the Metastases study (see Figures 1 and 2)

[d]Naïve unweighted analysis adjusted for assumed confounders.

Our work demonstrates that when case and control subjects are selected from an overlapping cohort, even with different sampling criteria, it is possible to reuse prior NCC data to address a new research question, provided information is available on the event times and matching factors that defined the sampling strategies for the prior study and the new cases, so that the common underlying study base can be defined. A naïve unweighted approach will provide biased estimates due to the available controls being unrepresentative of the appropriate control population as a result of the matching in the prior study. It is difficult to judge in advance the extent of the bias for the main exposure(s), but one could expect serious bias in the matching factors, as they contribute to the sampling and thus to the unrepresentative control data.[27] In contrast, the coefficients of the exposure variables from the weighted or unweighted analysis are both adjusted for the confounding by the matching factors so that any bias is a consequence of inadequate adjustment. Thus one might expect more serious bias in the coefficients where confounding effects are large. In our study, the risk factor estimates were only slightly biased while the estimates for age and adjuvant treatment (the matching factors in the prior study) were more seriously biased as expected. The low bias in the risk factor estimates is consistent with small confounding effects due to the matching factors. The lack of bias from using only the sampled controls was also highlighted in our study (Table 2) where there was no noticeable difference in the estimates.

The ability to reuse control data provides not only statistical efficiency but the potential for significant savings in cost and time. In our case, we saved the time and cost which would be needed to first request government authorities to conduct a data linkage to identify suitable controls, and then collect and record clinical details for 528 patients from their individual medical charts in the hospitals. The procedures for data linkage and access have strict and time-consuming protocols and the retrieval of clinical data is a laborious exercise, so it makes sense to avoid expending similar effort in new data collection if the existing data are appropriate to the research question at hand and can be validly reused. When the collection of the data involves biomarker measurement, the savings in reusing previously collected data are even more obvious, as in addition to the identification and recruitment of study participants, the costs associated with collection, pre-processing, storage and measurement of biological specimens are all avoided. Another advantage of our method is that the analysis handles the entire set of cases and available controls, whereas in traditional NCC studies, the conditional logistic regression cannot accommodate unmatched cases or controls, or case-control pairs that are concordant for exposure. Thus in some settings, an analysis that reuses control data can have better statistical power than the traditional approach. Efficiency gains from novel use of controls from NCC settings have also been reported for the estimation of population exposure prevalence.[28] However, in addition to its simplicity of implementation, conditional logistic regression offers another important advantage, as it does not need to model the matching variables, which can be a challenge with the weighted method when the matching variable is a highly stratified categorical factor.

In practice, the prior NCC study and the new cases will rarely be sampled from exactly the same underlying cohort, so that the reconstruction of the appropriate study base must be done with care. For example, it may not be possible to precisely define the common study base due to missing information on some exclusion criterion for the prior study. However, an imperfectly reconstructed study base should not be a problem if the underlying cohort is large, as the risk sets will be large at all event times (even within the matching strata), and the weights thus not sensitive to small changes. An imperfectly reconstructed study base could, however, be an issue in small cohorts. We compared the results of our weighted analysis with those obtained using the study base from which the metastases cases were sampled rather than the correct common underlying study base. Although this imperfect study base comprised 19% more patients, our estimates were essentially the same (Table 2) as both study bases were large and rather similar, providing large risk set sizes at any event time, even after stratification.

An important strength of the weighted method is that it can be implemented using simple commands available in standard statistical software: a straightforward Kaplan–Meier table, which requires no parametric assumptions, is used to identify the risk sets from which the weights are readily computed and a weighted Cox regression of cases and controls provides valid estimates. Since the role of the study base is only to provide the weights, the necessary information can be obtained from de-identified data that may even be grouped, which is a further benefit. In the situation where all data are in the same data set so that the sampled controls are identifiable in the cohort, our approach is still valid though one could also take advantage of a recently developed R package (*multipleNCC*) that can analyse such data.[29,30] We used the flexible R statistical software to perform all steps of the analysis. Other software such as SAS or Stata can also be used, as the data preparation involves only simple commands and the calculation of the weights is based on a Kaplan–Meier table which is provided by all standard statistical software. The manipulation of this table involves only basic operations to compute the weights used for the weighted Cox regression.

In our study, only categorical matching was used. In studies using caliper matching, one can still use the Kaplan–Meier type of weights with suitable categorical grouping, provided that the matching is not too close. For fine matching, other methods would be more appropriate, such as the so-called GLM- or GAM-weights.[21,27] However, such weights predict the sampling probability using models that require all cohort members to have a variable that indicates if they have been sampled for the previous NCC study. An advantage of the Kaplan–Meier type of weights is that this sampling indicator is not required to calculate the risk sets sizes so that a weighted analysis can be conducted even when data are anonymised, as in our situation.

The method presented in this paper has some limitations. We assume covariates to be measured at baseline, which in our case was the date of diagnosis of breast cancer. However, the NCC design is also valid when time-dependent exposures are involved but it is important that exposure is measured at the same time for a case and their matched controls.[31] Since the available controls from a prior NCC study will have exposure measured at different times than the cases in a new study, the weighted method would require extension to accommodate such data if there is a question of the stability of the measure over time. An important example is when the exposure is a biomarker that was determined by a laboratory analysis, where the biological samples of a case and the matched controls are analysed together to control for batch effects, as it has been shown that where there are strong batch effects the weighted method could lead to biased estimates.[32] Another limitation of the method is that if there is insufficient overlap of the cohorts from which the two studies are sampled, the common underlying cohort may not be large enough to have sufficient power, further reducing the precision of the estimates. In our case, the Metastases study was performed in a restricted study period (1997–2005), so although we had information for CBC patients from 1976, the alignment of our data resulted in a significant reduction in the number of CBC cases available for analysis. We included CBC cases from 1992 to improve statistical power, arguing that during this five-year extension (1992–1996), there were no major changes in BC diagnosis or treatment. This was supported by a sensitivity analysis of the 106 CBC cases and 528 available controls from 1997 which provided similar estimates but with an expected loss of power (data not shown). Thus our conclusions concern women who were diagnosed with their initial breast cancer between 1992 and 2005, although the available controls are from 1997 to 2005. For these conclusions to be valid, patients diagnosed with BC between 1992 and 1996 must be similar in all aspects to patients diagnosed after 1996. We would not recommend doing such a time-extension as a general rule, as it is based on assumptions that are context-specific and can lead to difficulties in interpreting the estimates.

Before implementing the method presented here, researchers should assess the feasibility and validity using the following considerations: (i) the cohort must be well defined, as must the sampling criteria for the prior NCC study and the new cases; (ii) for all members of the cohort, complete information must be available for a minimum set of variables (entry and censoring/event dates for both outcomes) and this basic information must also include the matching factors if the first study used matched sampling; (iii) for individuals included in the prior NCC study, their status should be known at the final date of the new study; (iv) the controls (for the prior NCC study) must have been randomly selected within the strata defined by the matching variables and, conditional on the complete covariates, any missing exposure information must not depend on the unobserved value, i.e. they must be "missing at random"; (v) the covariates of interest for the new outcome must not only be present in the data collected for the prior NCC study, but also defined/measured the same way. In addition, as the alignment of the data sets to common criteria reduces the number of subjects, there is a greater potential for advantage from reusing data from a NCC study with fewer exclusions, for example regarding age or study period. A prior study that sampled a larger number of controls per case will also provide more potential information for reuse, enabling more new cases to be studied. It should be noted that to achieve similar efficiency as a regular 1:m NCC study, the ratio of available controls to new cases may need to be $> m$, as these reused controls may be less informative.[13]

As NCC studies are traditionally analysed with conditional logistic regression, follow-up dates are not always routinely collected. Being aware of the potential to reuse their NCC data for other studies should prompt investigators to pay attention to important dates during data collection. In our example, if we did not have the last follow-up date for the sampled controls and had used the date of sampling as a proxy, this would influence both the weight calculation and the Cox analysis, resulting in biased estimates (data not shown).

Regarding the clinical findings of our study, our analysis enabled the identification of multifocality as a new risk factor for CBC. This finding is consistent with the literature where multifocality is associated with the higher-risk lobular histological type.[33,34] We also found that parity is protective against CBC and that positive family history and non-ductal histological type are associated with an increased risk, findings that are consistent with other published work.[23–26]

In summary, this paper illustrates how data from a NCC study can be reused to address a new research question, provided basic information is available for the underlying cohort and the sampling scheme, and the controls are appropriately weighted in the analysis. The proposed method can be used even in a situation where all data are anonymised. By making more efficient use of existing data resources, the availability of this method of analysis can result in significant savings in the cost and time required to gather new controls.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Swedish Cancer Society (Cancerfonden) through grant CAN2009/1175 and by the Australian National Health and Medical Research Council (NHMRC) grant no. 1108967.

References

1. Vandenbroucke JP and Pearce N. Case-control studies: basic concepts. *Int J Epidemiol* 2012; **41**: 1480–1489.
2. Borgan O and Samuelsen SO. Nested case-control and case-cohort studies. In: Klein JP, van Houwelingen HC, Ibrahim JG, et al. (eds) *Handbook of survival analysis*. Boca Raton, FL: Chapman & Hall/CRC, 2013, pp.343–367.
3. Borgan O and Samuelsen S. A review of cohort sampling designs. *Norsk Epidemiologi* 2003; **13**: 239–248.
4. van Rein N, Cannegieter SC, Rosendaal FR, et al. Suspected survivor bias in case-control studies: stratify on survival time and use a negative control. *J Clin Epidemiol* 2014; **67**: 232–235.
5. Ritchie K, Carrière I, Berr C, et al. The clinical picture of Alzheimer's disease in the decade before diagnosis: clinical and biomarker trajectories. *J Clin Psychiatr* 2016; (in press).

6. Hensgens MP, Dekkers OM, Demeulemeester A, et al. Diarrhoea in general practice: when should a Clostridium difficile infection be considered? Results of a nested case-control study. *Clin Microbiol Infect* 2014; **20**: O1067–O1074.
7. Concato J, Peduzzi P, Kamina A, et al. A nested case-control study of the effectiveness of screening for prostate cancer: research design. *J Clin Epidemiol* 2001; **54**: 558–564.
8. Jackson ML, Nelson JC, Weiss NS, et al. Influenza vaccination and risk of community-acquired pneumonia in immunocompetent elderly people: a population-based, nested case-control study. *Lancet* 2008; **372**: 398–405.
9. Langholz B and Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *Am J Epidemiol* 1990; **131**: 169–176.
10. Ernster VL. Nested case-controls studies. *Prev Med* 1994; **23**: 587–590.
11. 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**: 5991–6008.
12. 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**: 70–79.
13. 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**: 1291–1302.
14. Støer N and Samuelsen S. Comparison of estimators in nested case-control studies with multiple outcomes. *Lifetime Data Anal* 2012; **18**: 261–283.
15. Samuelsen S. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* 1997; **84**: 379–394.
16. Carpenter JR, Kenward MG and Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J R Statist Soc A* 2006; **169**: 571–584.
17. Seaman SR and White IR. Review of inverse probability weighting for dealing with missing data. *Stat Meth Med Res* 2013; **22**: 278–295.
18. Kim RS. A new comparison of nested case–control and case–cohort designs and methods. *Eur J Epidemiol* 2015; **30**: 197–207.
19. Samuelsen SO, Ånestad H and Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. *Scand J Stat* 2007; **34**: 103–119.
20. Salim A, Ma X, Li J, et al. A maximum likelihood method for secondary analysis of nested case-control data. *Stat Med* 2014; **33**: 1842–1852.
21. Støer N, Meyer HE and Samuelsen S. Reuse of controls in nested case-control studies. *Epidemiology* 2014; **25**: 315–317.
22. Vinogradova Y, Coupland C and 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**: e000548.
23. Bernstein JL, Lapinski RH, Thakore SS, et al. The descriptive epidemiology of second primary breast cancer. *Epidemiology* 2003; **14**: 552–558.
24. Vaittinen P and Hemminski K. Risk factors and age-incidence relationship for contralateral breast cancer. *Int J Cancer* 2000; **88**: 998–1002.
25. Reeves GK, Pirie K, Green J, et al. Reproductive factors and specific histological types of breast cancer: prospective study and meta-analysis. *Br J Cancer* 2009; **100**: 538–544.
26. Schaapveld M, Visser O, Louwman WJ, 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**: 189–197.
27. Støer N and Samuelsen S. Inverse probability weighting in nested case-control studies with additional matching – a simulation study. *Stat Med* 2013; **32**: 5328–5339.
28. Saarela O and Hanley JA. Case-base methods for studying vaccination safety. *Biometrics* 2015; **71**: 42–52.
29. Støer N and Samuelsen SO. Package ‘multipleNCC’. https://cran.r-project.org/web/packages/multipleNCC/multipleNCC.pdf (2015, accessed 9 September 2016).
30. Stoer NC and Samuelsen SO. MultipleNCC: inverse probability weighting of nested case-control data. *R Journal*, (in press).
31. Essebag V, Platt RW, Abrahamovicz M, et al. Comparison of nested case-control and survival analysis methodologies for analysis of time-dependent exposure. *BMC Med Res Methodol* 2005; **5**: 5.
32. Borgan O and Keogh R. Nested case-control studies: should one break the matching? *Lifetime Data Anal* 2015; **21**: 517–541.
33. 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**: 2551–2560.
34. Dedes KJ and Fink D. Clinical presentation and surgical management of invasive lobular carcinoma of the breast. *Breast Dis* 2008; **30**: 31–37.

Appendix 1. Computation of the Kaplan–Meier type weights

Assuming a sampled control i was followed up from a starting time s_i until time T_i and was eligible as a potential control at all event times during this interval, the probability (p_i) for i to be sampled at least once during the study depends on the probability of selection at each event time between s_i and T_i, which in turn depends on the risk set size (R_j) at each event time (T_j), and on the number of controls (m_j) sampled at T_j. The probability of not being sampled at all[13,15] is:

$$(1 - p_i) = \Pi_{\mathrm{j},\, si \leq Tj \leq Ti}\left[1 - m_j/\left(R_j - 1\right)\right] \tag{1}$$

where the product is taken over all events (i.e. cases) j occurring at times T_j.

In the case of stratification, the risk set sizes and thus p_i need to be computed as above in each of the strata.

Thus, to calculate the weights for reused controls from studies with or without matching, the required steps are as follows:

No matching

1. In the appropriate study base, define the outcome for which the controls were originally sampled (i.e. the prior outcome).
2. Generate a Kaplan–Meier table. In R, this is obtained by running the *survfit* command on the *Surv* object created from the data in 1.
3. The Kaplan–Meier table output provides the sizes of the risk sets (R_j). In R, *summary (survfit)* is a "list" that can be converted into a two-column data frame to enable processing of the two variables required for the weights (event times T_j and risk set sizes R_j). This data frame is expanded with the additional columns computed at each of the following steps.
4. For each T_j, calculate $(1 - m_j/(R_j - 1))$, with m_j the number of controls selected at T_j.
5. For each event time, compute the probability of not being sampled using the cumulative product in equation (1). This is run in R with the *cumprod* command.
6. Subtract the probability in step 5 from 1 to give the probability of being sampled.
7. For each control i, identify the last event time T_j for which he/she was at risk and assign the weight by taking the reciprocal of the probability corresponding to j.

With matching variables

Where the original nested case-control study used matching factors, a stratified Kaplan–Meier table is generated in step 2, and all operations in the subsequent steps are performed within each stratum, with step 3 requiring additional lines of code to transform the "list" into a data frame that includes the stratum identifier in an additional column. The cumulative product in step 5 can be computed within each stratum using the command *ddplyr* from the *plyr* package.