Breaking the matching in nested case—control data offered several advantages for risk estimation

Bénédicte Delcoigne a,*, Edoardo Colzani a, Michaela Prochazka a, Giovanna Gagliardi b, Per Hall b, Michal Abrahamowicz c, Kamila Czene a, Marie Reilly a

a Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, P.O. Box 281, SE-17177 Stockholm, Sweden
b Section of Radiotherapy Physics and Engineering, Department of Medical Physics, Karolinska University Hospital, P.O. Box 260, Stockholm SE-17176, Sweden
c Department of Epidemiology and Biostatistics and Occupational Health, McGill University, Purvis Hall, 1020 Pine Avenue West, Montreal, Quebec H3A 1A1, Canada

Accepted 28 November 2016; Published online 5 December 2016

Abstract

Objective: To demonstrate the advantage of using weighted Cox regression to analyze nested case—control data in overcoming limitations encountered with traditional conditional logistic regression.

Study Design and Setting: We analyzed data from 1,051 women who were sampled in a case—control study of lung cancer nested within a cohort of breast cancer patients. We investigated how lung cancer risk is associated with radiation therapy and modified by smoking, with both conditional logistic regression and weighted Cox regression models.

Results: In contrast to logistic regression, weighted Cox regression exploited the information regarding radiation dose received by each individual lung. The weighted method also mitigated a problem of overmatching apparent in the data and revealed that the risk of radiotherapy-associated lung cancer was modified by smoking (P = 0.026) with a hazard ratio of 4.09 (2.31, 7.24) in unexposed smokers and 8.63 (5.04, 14.79) in smokers receiving doses ≥ 13 Gy. The cumulative risk of lung cancer increased steadily with increasing radiotherapy dose in smokers, whereas no such effect was found in nonsmokers.

Conclusion: The weighted Cox regression makes optimal and versatile use of the information in a nested case—control design, allowing dose—response analysis of exposure to paired organs and enabling the estimation of cumulative risk. © 2016 Elsevier Inc. All rights reserved.

Keywords: Efficiency; Inverse probability weighting; Weighted Cox regression; Weighted likelihood; Absolute risk; Cumulative incidence

1. Background and motivation

The nested case—control design aims to combine the advantages of the cohort and case—control designs, namely, to benefit from the time dimension included in the outcome or exposure measurement in the cohort design while enabling significant savings in cost and time [1–4]. In the nested case—control design, complete exposure data are collected for all cases but only for a random sample of controls.

Using a risk set sampling strategy, a prespecified number of controls are sampled at each failure time among all individuals who are at risk at that time [2,3,5]. The traditional way of analyzing such data is to perform a conditional logistic regression, stratified by matched case—control sets. Although this method is easily implemented in standard statistical software, it has some limitations: (1) as a consequence of the matching, the matching factors cannot be easily studied as risk factors [6]; (2) case—control sets that are concordant for exposure do not contribute to the analysis and overmatching at the design stage can reduce the statistical power [7,8]; (3) although the matching on time enables hazard ratios to be estimated from a multivariable conditional logistic regression model, the cumulative incidence rate cannot be estimated due to the “matching away” of time in the risk sets, unless the time information is recovered [9–11]; and (4) when the research question concerns paired organs (e.g., lung, breast, eye) for each of which...
What is new?

Key findings
- Breaking the matching of nested case-control data and analyzing with weighted Cox regression can overcome several limitations encountered in conditional logistic regression.
- In an application to lung cancer risk following radiation for breast cancer, the weighted method mitigated a problem of overmatching at the design stage, accommodated radiation doses to the two lungs, and enabled estimation of the absolute risk of lung cancer for different radiation doses.
- The cumulative incidence of the outcome can be estimated due to the recovery of the time information.

What this adds to what was known?
- By using the weighted method, we demonstrated an interaction between radiotherapy and smoking in breast cancer patients, and a dose-response effect in smokers.

What is the implication and what should change now?
- The weighted method should be considered as an alternative to conditional logistic regression as it makes optimal and versatile use of all available information from all sampled subjects (including case-control sets concordant for exposure and information from paired organs).

Because the data described previously had been sampled from a well-defined cohort (i.e., a national register), the weighted partial likelihood method is an appropriate method of analysis [24]. The objective of this study is to
demonstrate the advantages of using weighted partial likelihood to optimize the statistical power for investigating a potential interaction between radiation and smoking status by including all information collected from all the sampled study subjects and also widen the possibilities for analyses, including the estimation of absolute risk \([10,11]\) and how this risk changes with dose to the lung in smokers and nonsmokers.

2. Data source and study design

One hundred sixty-four thousand two hundred twenty-eight (164,228) breast cancer patients were recorded in the nationwide Swedish Cancer Register \([25]\) between 1958 and 2001. This register can be considered as a cohort in which the patients were followed up for lung cancer after breast cancer diagnosis, with a median and maximum follow-up time of 18 and 44 years, respectively. Nine hundred (900) women were identified with a primary lung cancer subsequent to their breast cancer diagnosis. The censoring times used in this study were date of death, last follow-up or 31 December, 2001, whichever occurred first. Patients diagnosed with a second breast cancer during follow-up were censored at the date of this second breast cancer diagnosis if they received radiotherapy at that time, which was the case for \(4\%\) of the patients. Among the lung cancer patients, 730 women had medical charts available with dates of both diagnoses. One hundred ninety-three (193) lung cancer cases from the Stockholm-Gotland region were available for the original case-only analysis before data were collected from medical charts for the rest of Sweden.

\[\text{Cases diagnosed with LC at least 5 years after BC diagnosis and their matched controls:} \ n=1134 \ (565 \text{ cases} / 565 \text{ controls} + 4 \text{ unmatched cases})\]

\[\text{Patients with information on treatment received for breast cancer} \ n=1063 \ (509 \text{ sets with 1 case/1 control, 29 cases and 16 controls})\]

\[\text{Patients available for analysis} \ n=1051 \ (538 \text{ cases} / 513 \text{ controls})\]

---

**Fig. 1.** Flowchart outlining the selection of the study sample. \(^a\)The 193 patients were considered in a case-only design \([21]\). BC, breast cancer; LC, lung cancer.
of Sweden (see Fig. 1) [21]. Using incidence density sampling, controls were sampled within the same register. Controls had to be free of lung cancer at the case-defining date and to match their case on three additional variables: the decade of breast cancer diagnosis, age at breast cancer diagnosis ± 5 years and the region of residence (Stockholm-Gotland or the rest of Sweden).

We restrict our current analysis to cases whose lung cancer occurred at least 5 years after the breast cancer diagnosis to facilitate comparison with published studies [19–23], as 5 years is estimated as a reasonable latency period for observing a radiation-induced solid tumor [20,26,27]. This restriction resulted in 565 patients and their matched controls and four additional cases for whom no matched control was assigned. Among these 1,134 patients, 71 were missing information on the adjuvant treatment received for breast cancer (i.e., whether they had received radiotherapy or not) and were removed from the data set. Our final data set included 1,063 patients grouped in 509 matched case—control sets, and, due to the selection procedure mentioned previously, 29 unmatched cases, and 16 unmatched controls. As allowed by the nested case—control procedure, there were patients who had been sampled several times: four cases had been sampled as controls before their event time and eight controls had been sampled twice. For the analysis with weighted partial likelihood, we broke the matching of the 509 matched sets, and kept, as required by the weighted partial likelihood approach [28], all available patients with a unique record, each of them contributing in the analysis until their last follow-up time (event of lung cancer or censoring). Thus, the four cases who had been previously sampled as controls contributed only their case record, and the eight controls who had been sampled twice contributed just 1 record with their last follow-up time. The flowchart in Fig. 1 summarizes all these steps in the preparation of the final analysis data set consisting of 1,051 patients (538 cases and 513 controls).

In addition to the information on dates of diagnosis for both breast and lung cancers, the data include information about laterality of the cancers, information on smoking habits at the time of breast cancer diagnosis and details of the radiation therapy including radiation technique and estimated mean cumulative radiation dose to both the ipsilateral and contralateral (relative to breast cancer) lungs in Gray (Gy). The dose estimation, which has been described in detail previously, was conducted for each radiotherapy technique [21]. In our analysis, we categorized the doses in four categories: 0 Gy, <5 Gy, 5–13 Gy, and >13 Gy, that discriminated between the different techniques. In addition, women who did not have radiotherapy were assigned a dose of 0 Gy for both lungs. Regarding smoking habits, patients were categorized as smoker (including former smokers) or nonsmoker at the time of breast cancer diagnosis using information obtained from medical records and from questionnaires sent to the surviving patients or to the closest living relative [21,29]. Patients were followed for lung cancer starting from 5 years after breast cancer diagnosis and controls were censored at date of death, last follow-up or 31 December 2001, whichever occurred first. Patients diagnosed with a contralateral breast cancer during follow-up were censored at the date of diagnosis if they received radiotherapy for the contralateral breast cancer.

3. Statistical methods

We first compared the hazard ratio estimates obtained from the conditional logistic regression approach and weighted partial likelihood, with lung cancer (in either lung) as the outcome, and radiotherapy and smoking as binary exposures. We performed univariate analyses for each of these risk factors considered alone, and in the multivariable analysis included an additional multiplicative term to accommodate a potential interaction between them. In addition to the exposure variables of interest, the weighted analyses were adjusted for the matching factors, thus providing adjusted estimates that are comparable to those from the multivariable conditional logistic regression.

3.1. Weighting

To analyze the data using weighted likelihood, each control must be weighted by the inverse probability of being sampled for the study [24]. Each sampled control i is followed up from a starting time $s_i$ until his/her censoring time $T_i$ and is eligible as a potential control at all event times $T_j$ during this interval. The probability for i to be sampled at least once during the study depends on the probability of selection at each event time $T_j$ between $s_i$ and $T_i$, which in turn depends on the size of the risk set $R_j$ at this time. For a 1:1 case—control study with no matching other than time, this probability is given by $(1 - p_i) = \prod_{i \leq j \leq T_i} [1 - 1/(R_j - 1)]$ which reflects the incidence density sampling [12]. The product is taken over all event times at which the control is available for selection and gives the probability that individual i is not sampled at all. The sampling procedure in our study included matching variables in addition to time so that the risk set sizes and thus the $(1 - p_i)$ were computed as mentioned previously in each of the strata defined by the combinations of the matching factors.

To compute the sampling probabilities, basic information for all members of the underlying cohort is needed, that is, entry and censoring/event dates and the values of the matching variables. Because our data were sampled from a national register, this information was available. The risk set sizes were computed from a Kaplan–Meier table stratified on the three matching variables used for the sampling, that is, decade of breast cancer diagnosis, age in 10-year categories, and region of residence. All subsequent computation used simple algebra to arrive at the
sampling weights, which are known as Kaplan–Meier type weights [17]. By weighting the selected controls by the inverse of their sampling probability, they represent all the similar patients in the underlying cohort who might have been selected. In contrast, case patients get a weight of one to reflect the certainty of being selected. Once the weights are calculated, the data are readily analyzed using weighted Cox regression available in any standard statistical software package.

3.2. Dose–response analysis

As data were available on the mean cumulative radiation dose received at each lung, we performed a weighted Cox regression following each lung from 5 years after breast cancer diagnosis until the patient was diagnosed with lung cancer or censored as defined previously. The Cox model included these mean radiotherapy doses subdivided into four categories, binary smoking status, and their two-way interaction. We used the same Kaplan–Meier weights as in the previous analysis. The Cox model also included a cluster term to take into account the dependency of the outcomes for the two lungs in the same patient.

3.3. Absolute risk calculation

The absolute risk for a lung to develop cancer can be estimated from our model for any subgroup of interest. As an illustration, we calculated this risk at various time points from 5 to 25 years after breast cancer diagnosis for patients aged 54 years at breast cancer diagnosis (the mean age in our data), using the hazard ratio estimates from the weighted Cox regression analysis described in the previous paragraph, and the baseline hazard function estimate with the adapted Breslow estimator as proposed by Cai and Zheng [10] and illustrated in detail by Salim et al. [11].

3.4. Software

All data management and analysis were performed with the R statistical package (version 3.1.2) (http://www.R-project.org). To estimate the Kaplan–Meier weights and run the Cox regression analysis, we used the survfit and coxph functions, respectively, provided in the survival package. The baseline hazard was estimated by using the command basehaz with the option centered = FALSE.

4. Results

The univariate and adjusted estimates from the traditional conditional logistic regression analyses are presented in the two first left columns of Table 1. The coefficients [log(hazard ratio)] and their standard errors are presented in the first column, whereas the hazard ratios with the corresponding confidence intervals are in the second column. Smoking was identified as being significantly associated with lung cancer. There was an increased risk of developing lung cancer for women who received radiotherapy, but the result was not statistically significant. There was some evidence of an interaction between radiotherapy and smoking, but this did not achieve statistical significance, despite the use of a much larger data set than in the previous published work [21]. Including age as a continuous variable in the analysis to control for residual confounding did not alter the estimates nor the standard errors/confidence intervals presented in Table 1. The lack of statistical power for estimation of the radiotherapy effect became apparent from a descriptive analysis of the distribution of treatment within risk sets, where we found that in 286 of the 509 sets, both the case and control were treated with radiotherapy, and for 57 sets, neither case nor control received radiotherapy. Thus, 67% (343/509) of the sets are uninformative to study the radiotherapy exposure due to cases and controls being concordant for exposure. Examining the use of radiotherapy for breast cancer across calendar time, we found that before the mid 1970s, more than 80% of patients received radiotherapy as adjuvant treatment, so that the matching on calendar time had induced a high concordance in exposure in the matched sets (Fig. 2) with a consequent loss of power in the conditional analysis.

Breaking the matching, we analyzed the data using weighted Cox regression models adjusting for the matching factors. The median time to event was 14 years after breast cancer and the median time to censoring was 21 years. The estimates (coefficients and standard errors, hazard ratios, and confidence intervals) are presented in the two last columns of Table 1. The estimates are similar in magnitude to those from the conditional logistic regression, with smaller (or equal) standard errors for all analyses and narrower confidence intervals for the multivariate approach. We also obtained a significant interaction between smoking and radiotherapy.

The results from the weighted Cox regression analysis of individual lungs are presented in Table 2. In addition to the two exposures and their interaction term, we also included age at breast cancer diagnosis in the model, as the univariate analyses revealed that this was a potential confounder (data not shown). 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 nonsmokers. Inclusion of additional potential confounding factors, such as tumor stage at breast cancer diagnosis, use of chemotherapy and/or hormonal therapy and laterality of the breast cancer, did not substantially alter the hazard ratio estimates.

The estimates of absolute risk for a lung cancer at various time points from 5 to 25 years after breast cancer diagnosis are presented in Fig. 3 for women aged 54 years at breast cancer diagnosis. There is a clear trend in the
estimated risk over elapsed time and with radiation dose in smokers, the highest risk estimates being for the smokers receiving the highest radiation dose.

5. Discussion

In this study, we used weighted Cox regression models to analyze nested case–control data and illustrated how this nontraditional approach allowed for greater flexibility and a wider choice of analyses compared to conditional logistic regression. In an analysis of the risk of lung cancer subsequent to breast cancer, breaking the matching mitigated a problem of overmatching and improved the statistical power, revealing a significant interaction between smoking and radiotherapy. The weighted analysis also enabled straightforward handling of clustered data (i.e., pairs of lungs) so that detailed exposure information could be used which in turn facilitated the investigation of how the absolute risk of lung cancer depends on the received radiation dose to the lung and the smoking status of the patient.

Our findings of an increased relative risk of lung cancer with increasing radiation dose in smokers are consistent with the literature [21–23]. In contrast to Prochazka et al. who treated the lung on the contralateral side as unexposed (limiting analysis to techniques for which the received dose was negligible compared to the ipsilateral side), we took advantage of exploiting the values of the doses received by both the ipsilateral and contralateral lungs. The strength of the weighted method is thus that all the information on radiation dose to both lungs could be readily used, whereas the handling of paired organs in conditional logistic regression would require selection of a subsample for analysis (e.g., a case-only design as in Prochazka et al. that we mentioned previously [21] or cases and their ipsilateral controls [23]) with a consequent loss in power. The method that we used also enabled us to show

Table 1. Adjusted coefficients (log HR) with standard errors (SEs) and hazards ratio (HR) with 95% confidence interval (CI) for developing lung cancer 5 years or more after breast cancer

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Conditional logistic regression</th>
<th>Weighted Cox regressiona</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log HR (SE)</td>
<td>HR (95% CI)</td>
</tr>
<tr>
<td>Univariate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No radiotherapy</td>
<td>Ref</td>
<td>Ref</td>
</tr>
<tr>
<td>Radiotherapy</td>
<td>0.19 (0.16)</td>
<td>1.21 (0.89, 1.65)</td>
</tr>
<tr>
<td>No smoking</td>
<td>Ref</td>
<td>Ref</td>
</tr>
<tr>
<td>Smoking</td>
<td>1.73 (0.19)</td>
<td>5.64 (3.89, 8.16)</td>
</tr>
<tr>
<td>Multivariable + interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No radiotherapy and no smoking</td>
<td>Ref</td>
<td>Ref</td>
</tr>
<tr>
<td>Radiotherapy</td>
<td>–0.003 (0.29)</td>
<td>0.997 (0.56, 1.76)</td>
</tr>
<tr>
<td>Smoking</td>
<td>1.37 (0.32)</td>
<td>3.97 (2.13, 7.41)</td>
</tr>
<tr>
<td>Interaction</td>
<td>0.54 (0.39)</td>
<td>1.71 (0.80, 3.65)</td>
</tr>
</tbody>
</table>

The conditional logistic regression is performed with 1,018 patients in 509 matched sets, and the weighted Cox regression is performed with 1,051 unique patients.

a All weighted analyses include adjustment for the matching variables used in the sampling, that is, age (continuous), region, and decade of diagnosis.

Table 2. Hazard ratio (HR) and 95% confidence interval (CI) for a lung to develop cancer 5 years or more after breast cancer

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>HR (95% CI)</th>
<th>P for trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>No radiotherapy—nonsmoking</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>No radiotherapy—smoking</td>
<td>4.09 (2.31, 7.24)</td>
<td>0.026b</td>
</tr>
<tr>
<td>Radiotherapy—nonsmoking &lt;5 Gy</td>
<td>0.74 (0.42, 1.30)</td>
<td></td>
</tr>
<tr>
<td>5–13 Gy</td>
<td>0.92 (0.49, 1.75)</td>
<td></td>
</tr>
<tr>
<td>13 Gy+</td>
<td>0.77 (0.43, 1.38)</td>
<td></td>
</tr>
<tr>
<td>Radiotherapy—smoking &lt;5 Gy</td>
<td>5.42 (3.06, 9.60)</td>
<td></td>
</tr>
<tr>
<td>5–13 Gy</td>
<td>7.15 (3.60, 14.20)</td>
<td></td>
</tr>
<tr>
<td>13 Gy+</td>
<td>8.63 (5.04, 14.79)</td>
<td></td>
</tr>
</tbody>
</table>

The weighted Cox regression was performed for the 2,102 lungs of the 1,051 study patients, with a cluster term for patient.

a The analyses were adjusted for age as a linear term.

b The P for trend was calculated for smokers only.
that the risk of developing lung cancer is modified by the smoking status of the patient and that the absolute risk of lung cancer in smokers increases with the radiotherapy dose received at the lung site.

The data in this study were sampled following a nested case–control design that was characterized by two weaknesses which limited the performance of standard conditional logistic regression analysis. The relatively small number of controls limited the power of the analysis compared to studies with higher case–control ratios [22,23] and made the conditional analysis more vulnerable to loss of risk sets due to concordant exposure. This loss was further exacerbated by matching unnecessarily [30] on decade of treatment, which resulted in cases and controls being exposed to similar treatment protocols or doses of radiation. The loss of power due to these features was reflected in the larger standard errors of the estimated coefficients in the conditional logistic regression (Table 1) and, to a lesser extent, the inability of the conditional logistic regression analysis to identify the interaction between radiation and smoking. The highlighting of the interaction term in the weighted analysis was not only due to better efficiency (smaller standard errors) but also to the larger magnitude of the point estimate. To estimate the standard errors from the weighted Cox regression, we used a robust variance estimator that assumes the weights to be known. To check whether this assumption affects the validity of the standard errors, we recomputed the estimates using a bootstrap procedure and found very similar (slightly narrower) confidence intervals, indicating that the simple-to-use robust estimator is reasonable.

In the dose–response analyses, radiation doses were classified in four categories. Although this might be perceived as a loss of information, it enabled us to avoid any assumption of a linear relationship between dose and risk. Moreover, the categorization accommodates the potential inaccuracy in how the dose levels were reconstructed from information on radiotherapy technique [21]. In a supplementary analysis, we implemented the weighted Cox regression analysis with the dose as a continuous variable and obtained similar results: for example, the hazard ratio (95% confidence interval) for a smoker who received a radiotherapy dose of 15 Gy was 8.37 (5.61, 12.48).

In keeping with the literature on radiation-induced lung cancer [20,26,27], we used a 5-year latency period between breast cancer diagnosis and lung cancer diagnosis. Interestingly, the hazard ratio was below 1 within the first 5 years, consistent with the suggestion that radiotherapy during this period may offer a protective effect by eradicating any latent cancer which was developing at the time of breast cancer diagnosis [31]. However, when starting follow-up at 5 years after breast cancer diagnosis, there was still some evidence of nonproportional hazards for the radiation doses in four categories. The proportional hazards assumption was fully met when using a 10-year latency period—a latency suggested by some clinical researchers [21,22]—together with a finer level of categorization for the doses (i.e., 0, 0–5, 5–13, 13–21, and > 21 Gy), and the estimates obtained with these latency and dose categorizations were similar to those presented in Tables 1 and 2 but with some loss of power (data not shown).

A limitation of the Kaplan–Meier type of weights on our setting is that although age was used as a caliper matching variable (i.e., age ±5 years) in the sampling design, the weights were computed using 10-year age categories, a reasonable choice to avoid categories that were too narrow [17]. To assess the sensitivity of our results to these weights, we reanalyzed the data with GLM weights [17] and obtained very similar estimates (data not shown) indicating that the categorical age variable captures the sampling probabilities inherent in the caliper matching. For a weighted likelihood analysis of nested case–control data to be feasible and valid, complete information must be available for a minimum set of variables (entry and censoring/event dates and matching factors) for all members of the underlying cohort. In addition, the incomplete data from the incidence density subsampling of the cohort must be missing at random, which requires that controls have been randomly selected within the strata defined by the matching variables and that any missing exposure information does not depend on the unobserved value [2,13].

In addition to illustrating a useful methodology for more efficient use of nested case–control data, we obtained results of potential clinical interest, namely that the association of breast cancer irradiation on lung cancer risk is different (i.e., larger) for smokers and that this risk increases with increasing radiation dose. Although these findings are in line with other published work [21–23], they should be interpreted with caution. The absolute risk estimation treated death as a censoring event, so that we have estimated the risk of lung cancer that would be observed if there was no competing risk of death. In addition, as the...
confidence intervals were rather wide, the dose reconstruction procedure subject to inaccuracies, and information on smoking was a simple binary variable [29,32], further investigation is necessary to enable conclusions regarding clinical relevance.

In summary, this study demonstrates several advantages of weighted likelihood over conditional logistic regression for matched nested case-control studies. This nontraditional analysis makes more efficient use of all the available data, can overcome problems due to overmatching, and allows an extended choice of analyses such as estimating absolute risk and handling health outcomes involving paired organs.

References