Journal of Clinical Epidemiology 82 (2017) 79–86

Journal of Clinical Epidemiology

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[a], 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.

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

Funding: This research was supported by the Swedish Cancer Society (Cancerfonden) through the grant CAN2009/1175. This research was also partially supported by the Canadian Institutes of Health Research (CIHR) Drug Safety and Effectiveness Network grant TD3-137716 from the Canadian Network for Advanced Interdisciplinary Methods for comparative effectiveness research (CAN-AIM).

Conflict of interest: None.

* Corresponding author. Tel.: +46 (0)8 524 839 82; fax: +46 (0)8 31 11 01.

E-mail address: benedicte.delcoigne@ki.se (B. Delcoigne).

http://dx.doi.org/10.1016/j.jclinepi.2016.11.014
0895-4356/

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).

exposure measurements are available, conditional logistic regression cannot readily handle such clustered data.

Alternative methods have been developed for the analysis of nested case–control data in which the matching between the cases and their controls is broken and hazard ratios are estimated by maximizing weighted likelihood equations [12,13]. These weighted methods have been investigated for their performance in numerous simulation studies and illustrative data analyses in the medical literature where they have been compared to standard epidemiological designs [14–18]. A weighted Cox regression has been shown to have an increased efficiency in the analysis of nested case-control data compared to the traditional conditional logistic regression approach [3,14–18], and the possibility of overcoming the limitations outlined previously was also mentioned [3,15]. In this study, we used such nontraditional analyses to address a research question concerning the risk of lung cancer after radiotherapy for breast cancer, where data on paired organs (breast and lung) were gathered in a nested case–control design that was overmatched. The conditional logistic regression was thus underpowered for the investigation of an interaction. In addition to mitigating this problem of overmatching and handling the paired data, the weighted partial likelihood enabled the estimation of cumulative risk.

The effect of radiation therapy for breast cancer as a potential risk factor for subsequent lung cancer has been investigated in several studies for women who had postoperative breast cancer radiation treatment. An increased risk of lung cancer has been shown for at least 5 years after the adjuvant treatment and even decades later [19,20]. Smoking is the main risk factor for developing lung cancer, and the risk of developing lung cancer after radiotherapy was shown to be particularly increased among smokers but not statistically significant in nonsmokers [21–23]. The interaction between these two carcinogens is however not fully understood, and, to the best of our knowledge, there is no published work investigating how the absolute (cumulative) risk of lung cancer in smokers and nonsmokers depends on radiation dose.

To address these questions, we used data that were collected using a case–control design nested within a cohort of Swedish breast cancer patients. Cases were breast cancer patients subsequently diagnosed with lung cancer, and incidence density sampling was used to select matched controls from breast cancer patients without any subsequent cancer diagnosis before their date of selection. In addition to time since breast cancer diagnosis, controls were matched on age, region of residence, and calendar period of diagnosis. A subset of the data, including only radiated women from the Stockholm region who were cases of lung cancer, was previously analyzed in a case-only design where the two lungs of the same woman were compared [21]. The authors found a relative risk of 3.17 (1.66–6.06) for smokers to develop lung cancer on the ipsilateral (radiated) side 10 or more years after radiotherapy, in contrast to a relative risk of 0.9 (0.37–2.22) in nonsmokers. They also reported evidence of a dose–response trend in smokers but not in nonsmokers. Although these findings suggest an interaction between smoking and radiation, the design did not allow formal testing of such interaction because both lungs in each woman were assigned to either the smoker or nonsmoker category, and the wide confidence intervals, due to small sample size, did not allow for a meaningful upper bound on the association of lung cancer with radiation in nonsmokers. Furthermore, this analysis did not make use of any information from controls, nor from cases who were not treated with radiotherapy, yet information was available for all of these individuals regarding the laterality of the breast and lung cancers, as well as the radiation dose received by each lung.

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

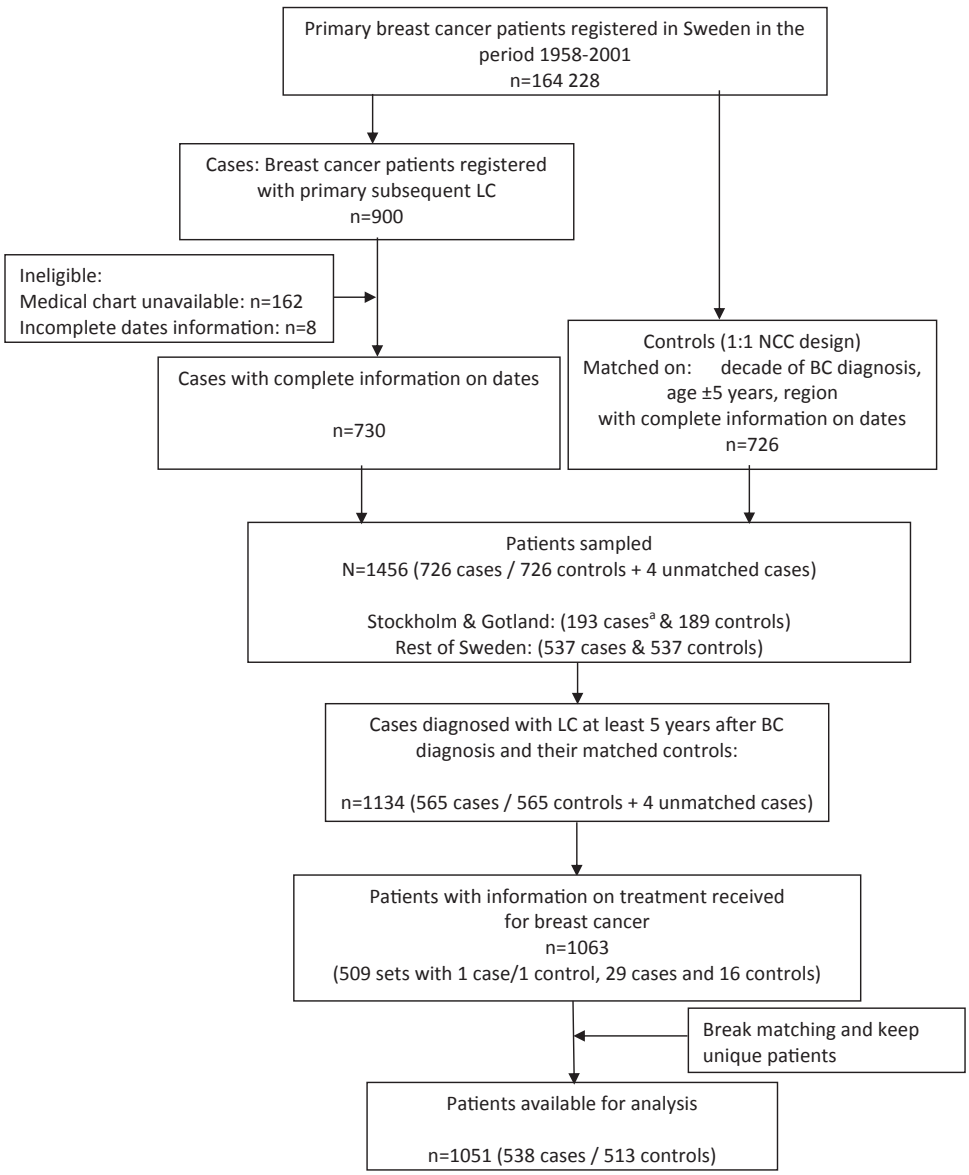

Fig. 1. Flowchart outlining the selection of the study sample. [a]The 193 patients were considered in a case-only design [21]. BC, breast cancer; LC, lung cancer.

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

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

	Conditional logistic regression		Weighted Cox regression[a]	
Risk factors	**Log HR (SE)**	**HR (95% CI)**	**Log HR (SE)**	**HR (95% CI)**
Univariate				
No radiotherapy	Ref	Ref	Ref	Ref
Radiotherapy	0.19 (0.16)	1.21 (0.89, 1.65)	0.20 (0.16)	1.23 (0.89, 1.69)
No smoking	Ref	Ref	Ref	Ref
Smoking	1.73 (0.19)	5.64 (3.89, 8.16)	1.94 (0.16)	6.93 (5.03, 9.58)
Multivariable + interaction				
No radiotherapy and no smoking	Ref	Ref	Ref	Ref
Radiotherapy	−0.003 (0.29)	0.997 (0.56, 1.76)	−0.26 (0.25)	0.77 (0.47, 1.26)
Smoking	1.37 (0.32)	3.97 (2.13, 7.41)	1.38 (0.29)	3.96 (2.25, 6.97)
Interaction	0.54 (0.39)	1.71 (0.80, 3.65)	0.78 (0.34)	2.19 (1.12, 4.30)

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.

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.

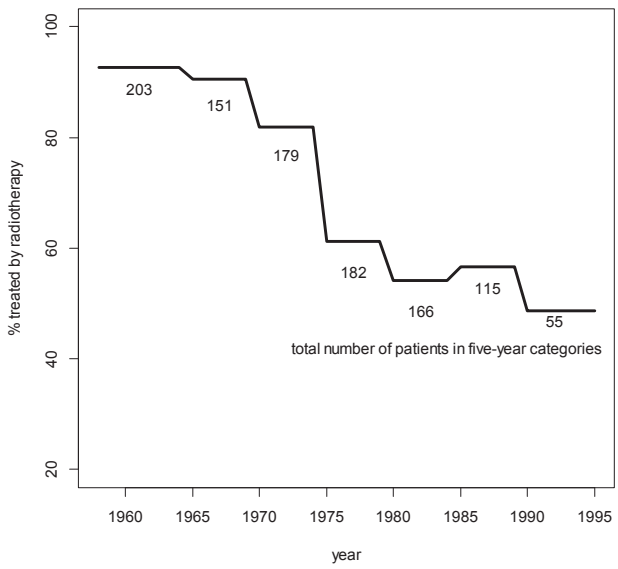

Fig. 2. Proportion of the 1,051 study patients who were treated by radiotherapy, as a function of calendar time and with the numbers of patients displayed on the plot.

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

Risk factors	**HR (95% CI)[a]**	***P* for trend**
No radiotherapy–nonsmoking	1	
No radiotherapy–smoking	4.09 (2.31, 7.24)	
Radiotherapy–nonsmoking		
<5 Gy	0.74 (0.42, 1.30)	
5–13 Gy	0.92 (0.49, 1.75)	
13 Gy+	0.77 (0.43, 1.38)	
Radiotherapy–smoking		
<5 Gy	5.42 (3.06, 9.60)	
5–13 Gy	7.15 (3.60, 14.20)	0.026[b]
13 Gy+	8.63 (5.04, 14.79)	

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.

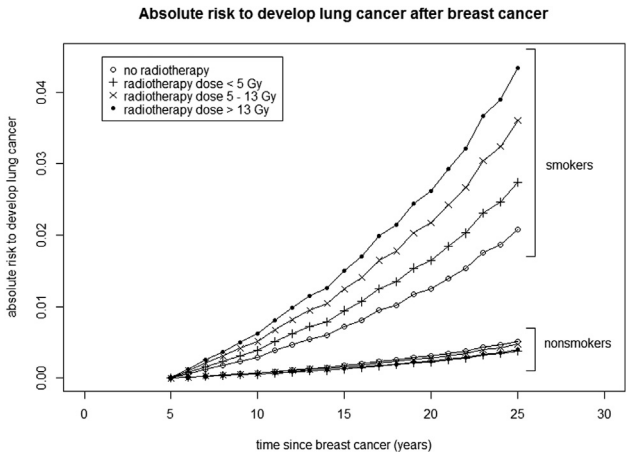

Fig. 3. The estimated absolute risk (i.e., probability) of cancer in a lung exposed to various radiation doses, estimated at various time points from 5 to 25 years after breast cancer for patients aged 54 years at breast cancer diagnosis, assuming no competing risk of death.

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

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

[2] Borgan O, Samuelsen SO. A review of cohort sampling designs for Cox's regression model: potentials in epidemiology. Nor Epidemiol 2003;13(2):239−48.

[3] Borgan O, Samuelsen SO. Nested case-control and case-cohort studies. In: Klein JP, van Houwelingen HC, Ibrahim JG, Scheike TH, editors. Handbook of Survival Analysis. Boca Raton, USA: Chapman & Hall/CRC; 2013:343−67.

[4] Essebag V, Platt RW, Abrahamovicz M, Pilote L. Comparison of nested case-control and survival analysis methodologies for analysis of time-dependent exposure. BMC Med Res Methodol 2005;5:5.

[5] Vandenbroucke JP, Pearce N. Case-control studies: basic concepts. Int J Epidemiol 2012;41:1480−9.

[6] Siskind V, Kelly JP, Kaufman DW. Estimating risks for matching factors in case-control studies. J Clin Epidemiol 2000;53:251−6.

[7] Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Chapter 11 & 16. 3rd ed. Philadelphia USA: Lippincott, Williams & Wilkins; 2008.

[8] Breslow NE, Day NE. The analysis of case-control studies. Lyon, France: IARC; 1980. Available at http://www.iarc.fr/en/publications/pdfs-online/stat/sp32/SP32_vol1-0.pdf. Accessed October 21, 2016.

[9] Langholz B, Borgan Ø. Estimation of absolute risk from nested case-control data. Biometrika 1997;53:767−74.

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

[11] Salim A, Delcoigne B, Villaflores, Koh WP, Yuan JM, van Dam RM, et al. Comparisons of risk prediction methods using nested case-control data. Stat Med 2016; http://dx.doi.org/10.1002/sim.7143.

[12] Samuelsen SO. A pseudolikelihood approach to analysis of nested case-control studies. Biometrika 1997;84(2):379−94.

[13] Saarela O, Kulathinal S, Arjas E, Läärä E. Nested case-control data utilized for multiple outcomes: a likelihood approach and alternatives. Stat Med 2008;27:5991−6008.

[14] Leffondre K, Wynant W, Cao Z, Abrahamowicz M, Heinze G, Siemiatycki J. A weighted Cox model for modelling time-dependent exposures in the analysis of case-control studies. Stat Med 2010;29:839−50.

[15] Samuelsen SO, Ånestad H, Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. Scand J Stat 2007;34(1):103−19.

[16] Kim RS. A new comparison of nested case−control and case−cohort designs and methods. Eur J Epidemiol 2015;30:197−207.

[17] Støer N, Samuelsen S. Comparison of estimators in nested case-control studies with multiple outcomes. Lifetime Data Anal 2012;18(3):261−83.

[18] Borgan Ø, Keogh R. Nested case−control studies: should one break the matching? Lifetime Data Anal 2015;21(4):517−41.

[19] 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):224−33.

[20] 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;114:56−65.

[21] 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;23:7467−74.

[22] 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;26:392−8.

[23] Grantzau T, Thomsen MS, Væth M, Overgaard J. Risk of second primary lung cancer in women after radiotherapy for breast cancer. Radiother Oncol 2014;111:366−73.

[24] Salim A, Hultman C, Sparén P, Reilly M. Combining data from 2 nested case-control studies of overlapping cohorts to improve efficiency. Biostatistics 2009;10(1):70−9.

[25] 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):305−13.

[26] The National Academy of Sciences (NAS) Biological Effects of Ionizing Radiation (BEIR). Health risks from exposure to low levels of ionizing radiation: BEIR VII − phase 2 (free executive summary). 2006. Available at http://www.nap.edu/catalog/11340.html. Accessed December 22, 2016. ISBN: 978-0-309-09156-5, 424 pages, 8 1/2 x 11, paperback (2006).

[27] Maddams J, Parkin DM, Darby SC. The cancer burden in the United Kingdom in 2007 due to radiotherapy. Int J Cancer 2011;129:2885−93.

[28] Salim A, Yang Q, Reilly M. The value of reusing prior nested case−control data in new studies with different outcome. Stat Med 2012;31:1291−302.

[29] Prochazka M, Hall P, Granath F, Czene K. Validation of smoking history in cancer patients. Acta Oncol 2008;47:1004−8.

[30] Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 2009;20:488−95.

[31] Prochazka M, Granath F, Ekbom A, Shields PG, Hall P. Lung cancer risks in women with previous breast cancer. Eur J Cancer 2002;38:1520−5.

[32] Rachet B, Siemiatycki J, Abrahamowicz M, Leffondré K. A flexible modeling approach to estimating the component effects of smoking behavior on lung cancer. J Clin Epidemiol 2004;57:1076−85.