Incorporating survival data into case-control studies with incident and prevalent cases

Soutrik Mandal1 | Jing Qin2 | Ruth M. Pfeiffer1

1Biostatistics Branch, National Cancer Institute, NIH, Bethesda, Maryland, USA
2Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, Maryland, USA

Correspondence
Ruth M. Pfeiffer, Biostatistics Branch, National Cancer Institute, NIH, Bethesda, MD, USA.
Email: pfeiffer@mail.nih.gov

Typically, case-control studies to estimate odds-ratios associating risk factors with disease incidence only include newly diagnosed cases. Recently proposed methods allow incorporating information on prevalent cases, individuals who survived from disease diagnosis to sampling, into cross-sectionally sampled case-control studies under parametric assumptions for the survival time after diagnosis. Here we propose and study methods to additionally use prospectively observed survival times from prevalent and incident cases to adjust logistic models for the time between diagnosis and sampling, the backward time, for prevalent cases. This adjustment yields unbiased odds-ratio estimates from case-control studies that include prevalent cases. We propose a computationally simple two-step generalized method-of-moments estimation procedure. First, we estimate the survival distribution assuming a semiparametric Cox model using an expectation-maximization algorithm that yields fully efficient estimates and accommodates left truncation for prevalent cases and right censoring. Then, we use the estimated survival distribution in an extension of the logistic model to three groups (controls, incident, and prevalent cases), to adjust for the survival bias in prevalent cases. In simulations, under modest amounts of censoring, odds-ratios from the two-step procedure were equally efficient as those estimated from a joint logistic and survival data likelihood under parametric assumptions. This indicates that utilizing the cases’ prospective survival data lessens model dependencies and improves precision of association estimates for case-control studies with prevalent cases. We illustrate the methods by estimating associations between single nucleotide polymorphisms and breast cancer risk using controls, and incident and prevalent cases sampled from the US Radiologic Technologists Study cohort.

KEYWORDS
exponential tilting model, left truncation, length biased sampling, survival bias

1 INTRODUCTION

Case-control studies are economical and therefore popular for estimating the association of exposures with disease incidence for rare outcomes. They typically include only incident cases, that is, individuals with newly diagnosed disease.
However, sometimes subjects who developed the disease before the start of the study, termed “prevalent cases,” are also available for sampling. Simply combining information from incident and prevalent cases leads to biased estimates of disease-exposure association, if the exposure of interest for disease incidence also impacts survival after disease diagnosis, as prevalent cases had to survive long enough to be available for sampling into the case-control study, that is, their observations are left truncated.

Maziarz et al. proposed a fully efficient method to estimate log-odds ratios of disease-exposure associations when prevalent cases are included in a cross-sectionally sampled case-control study. They extended the exponential tilting (or density ratio) model of Qin for the disease-exposure relationship to accommodate prevalent cases by correcting for their survival bias through a tilting term that depends on the distribution of the survival time following disease onset. However, for cross-sectionally sampled prevalent cases, only the backward times, defined as the times from disease diagnosis to sampling, are observed and thus a fully parametric model for the distribution of the survival time following disease onset is required for identifiability.

Often prospective follow-up information on the time from disease diagnosis to death for incident and prevalent cases is available. In this article, we propose and study novel methods to utilize this prospective information to relax parametric assumptions and to estimate the tilting term in the logistic model for prevalent cases based on the popular semiparametric Cox proportional hazards model. To our knowledge this is the first article to incorporate prospective survival information into the analysis of case-control studies with prevalent cases. Information on survival after disease onset can typically be obtained readily when cases and controls are sampled from a well-defined cohort, and investigators periodically update the vital status of cohort members by linking to national databases such as the US National Death Index. This is the setting of the study that motivated our work, a case-control study that included controls, and incident and prevalent breast cancer cases sampled from within the US Radiologic Technologists Study (USRTS) cohort to estimate the association between selected single nucleotide polymorphisms (SNPs) and breast cancer risk. A second important setting where prospective survival information is commonly available is for case-control studies that use data extracted from health insurance claims databases and electronic medical records. A related example is the widely used Surveillance, Epidemiology and End Results (SEER)-Medicare database, created by linking SEER cancer registries with Medicare claims from the time of a person's Medicare eligibility until death from which cancer cases can be sampled. The SEER-Medicare database also contains a 5% random sample of Medicare beneficiaries from which controls can be obtained (https://healthcaredelivery.cancer.gov/seermedicare/).

In this article, we derive a profile likelihood that combines the retrospectively sampled case-control data and the prospective survival data for incident and prevalent cases, while accommodating left truncation for the prevalent cases and right censoring. When the survival model is parametrically specified, we jointly maximize this likelihood for the logistic model and the Cox proportional hazards model. However, for a semiparametrically specified survival distribution the joint estimation is computationally prohibitive complex. We thus develop a two-step generalized methods of moments procedure to estimate log-odds ratios for association of exposures with disease incidence from case-control studies that also include prevalent cases when prospective survival data on cases are available. In the first step of the procedure we estimate the survival distribution based on a Cox proportional hazards model using prospective follow-up time information available on the cases. We propose an expectation-maximization (EM) algorithm that builds on work by Liu et al. and Qin et al. to obtain maximum likelihood estimates (MLEs) of the Cox model parameters for right-censored, left-truncated data that are fully efficient. We also compare these results with those from maximizing the Cox partial likelihood accommodating left truncation of the prevalent cases and right censoring in the data, that can be implemented using standard statistical software. In the second step, we use the estimated survival distribution in the semiparametric profile likelihood for controls, incident and prevalent cases and estimate log-odds ratios for disease-exposure associations. Section 3 details the two-step procedure after introducing the model and notation in Section 2.

We derive the asymptotic properties and study the performance of the proposed method using simulations with varying sample sizes and amounts of censoring of the prospective survival data (Section 4). We compare the small sample properties and the efficiency of log-odds ratio estimates that utilize prospective survival information with those of estimates from Maziarz et al. These results can help decide whether collecting additional prospective information on the incident and prevalent cases warrants the cost for any additional data collection. We illustrate the methods by analyzing the USRTS breast cancer case-control study, that included incident and prevalent cases (Section 5), before closing with a discussion (Section 6).
2  MODELS AND LIKELIHOOD

First we summarize the data and models for case-control studies with prevalent cases used in Maziarz et al\(^2\) and then incorporate prospective information on the time from disease diagnosis to death for the sampled cases.

2.1  Background: Semiparametric model for case-control studies with incident and prevalent cases

Let \(D\) denote the disease indicator, with \(D = 1\) for those with disease (cases) and \(D = 0\) for those without (controls), and let \(X\) denote a vector of covariates.

2.1.1  Exponential tilting models

We assume in the population \(X\) is associated with incident disease through the logistic model

\[
P(D = 1|X = x) = \frac{\exp(a_0 + x^T \beta)}{1 + \exp(a_0 + x^T \beta)},
\]

where \(a_0\) is the intercept term and \(\beta\) is the vector of log-odds ratios.

The marginal probability of disease in the population is \(\pi = P(D = 1) = \int P(D = 1|x)f(x)dx\) where \(f(x) = dF(x)/dx\) is the unspecified density corresponding to the cumulative distribution function \(F(x)\) of \(X\). In retrospectively sampled case-control studies we observe the conditional densities \(f_0(x) = f(x|D = 0)\) for controls, and \(f_1(x) = f(x|D = 1)\) for incident cases. Under the population model (1), these two densities are related through the exponential tilting (or density ratio) model\(^3\)

\[
f_1(x) = \frac{\exp(a_0 + x^T \beta)}{1 + \exp(a_0 + x^T \beta)} f(x)\pi = f_0(x) \exp(\nu^* + x^T \beta),
\]

where \(\nu^* = a_0 + \log((1 - \pi)/\pi)\).

Maziarz et al\(^2\) extended model (2) to accommodate covariates \(X\) from prevalent cases. However, only those prevalent cases are observed whose backward time \(A\), that is, the time between disease diagnosis and sampling into the case-control study, is shorter than the time \(T\) from diagnosis to death, that is, they are subjected to left truncation. Thus the joint distribution of the observed data \((X, A|A < T, D = 1)\) for prevalent cases is

\[
f(X = x, A = a|D = 1, T > A) = f(X = x|D = 1, T > A)f(A = a|X = x, D = 1, T > A).
\]

If the disease incidence is stationary over time, the backward time \(A\) has a uniform distribution in some interval \((0, \xi]\) (see detailed explanation in Maziarz et al\(^2\) and our comment on this assumption in the Discussion). Further assuming that the time to disease onset and \(T\) are independent, the density of \(X\) for prevalent cases is

\[
f_2(x) = f(x|D = 1, T > A) = f_0(x) \exp(\nu^* + x^T \beta + \log \mu(x, \kappa)),
\]

where \(S(t|x, \kappa) = P(T > t|x, \kappa)\) denotes the survival distribution of \(T\) with parameters \(\kappa\),

\[
\mu(x, \kappa) = \int_0^\xi S(a|x, \kappa)da,
\]

\(\nu^* = \alpha^* - \log(\int_\kappa^\xi \mu(x, \kappa)f_1(x)dx\) and \(\kappa\) is the support of \(X\). As \(A\) is independent of \(X\) and \(T\), the conditional density of \(A\), for \(a \in (0, \xi]\), is

\[
f_A(a|X = x, D = 1, T > A) = \frac{f(A = a)P(T > a|X = x, D = 1)}{P(T > A|X = x, D = 1)} = \frac{S(a|x, \kappa)}{\mu(x, \kappa)},
\]

and 0 for \(a \not\in (0, \xi]\). If only the backward times \(A\) are observed, but not the actual survival times, \(S\) needs to be specified fully parametrically to ensure identifiability.
Incorporating prospective follow-up information on the cases begins, \([T_{dr}, T_{dr}+\Delta]\) is the sampling period, \(T_{ds}\) is the diagnosis time for prevalent cases and \(T\) is the time from disease diagnosis to death for cases.

2.1.2 Profile log-likelihood

Using the exponential tilting models (2) and (3), and the backward time distribution (5), the likelihood for the cross-sectionally observed data for the controls and the two case groups is

\[
\mathcal{L} = \left\{ \prod_{i=1}^{N} f_0(x_i) \right\} \left\{ \prod_{i=n_0+1}^{n_0+n_1} \exp(\alpha^* + x_i^T \beta) \right\} \left\{ \prod_{i=n_0+n_1+1}^{N} \exp(\nu^* + x_i^T \beta + \log \mu(x_i, \kappa)) \right\} S(a_i|x_i, \kappa),
\]

where \((x_1, \ldots, x_{n_0})^T\) are the covariates for the \(n_0\) controls, \((x_{n_0+1}, \ldots, x_{n_0+n_1})^T\) the covariates for the \(n_1\) incident cases, and \((a_{n_0+n_1+1}, \ldots, a_N)^T\) the covariates and backward times for the \(n_2\) prevalent cases, with \(N = n_0 + n_1 + n_2\). To ensure that \(f_i, i = 0, 1, 2\) are distributions, the \(p_i = f_0(x_i) = P(X = x_i), i = 1, \ldots, N,\) are estimated empirically under the following constraints: \(\sum_{i=1}^{N} p_i = 1, p_i \geq 0; \sum_{i=1}^{N} p_i \exp(\alpha^* + x_i^T \beta) = 1;\) and \(\sum_{i=1}^{N} p_i \exp(\nu^* + x_i^T \beta + \log \mu(x_i, \kappa)) = 1\) via Lagrange multipliers. After maximizing the log-likelihood for \(p_i\) subject to the constraints, the profile log-likelihood for the remaining parameters \((\alpha, \nu, \beta, \kappa)\) is

\[
l_p(\alpha, \nu, \beta, \kappa) = -\sum_{i=1}^{N} \log[1 + \exp(\alpha + x_i^T \beta) + \exp(\nu + x_i^T \beta + \log \mu(x_i, \kappa))] + \sum_{i=n_0+1}^{n_0+n_1} (\alpha + x_i^T \beta) + \sum_{i=n_0+n_1+1}^{N} (\nu + x_i^T \beta + \log \mu(x_i, \kappa) + \log \left\{ \frac{S(a_i|x_i, \kappa)}{\mu(x_i, \kappa)} \right\},
\]

where \(\alpha = \alpha^* + \log(n_1/n_0)\) and \(\nu = \nu^* + \log(n_2/n_0)\).

2.2 Incorporating prospective follow-up information on the cases

We now assume that in addition to case-control status, covariates and the backward times for prevalent cases, prospective follow-up information on the time \(T\) from disease diagnosis to death is observed on all cases in the case-control study. This is the setting of our motivating study that sampled incident and prevalent breast cancer cases and controls from the USRTS cohort to assess the association of breast cancer risk with SNPs in select genes. Investigators regularly link USRTS cohort members with the US national death index (NDI), to update vital status data.

Letting \(C\) denote the censoring time, we define the observed prospective follow-up time to be \(Y = \min(T, C)\) and the event indicator \(\delta = 1\), if \(Y = T\) and \(\delta = 0\) otherwise. The data observed for an incident case are \(O = (Y, \delta, X = x)\) and for a prevalent case \(O = (Y, \delta, X = x|A < T),\) that is, the survival times are left-truncated in addition to being right-censored. We assume that \((T, A)\) and \(C\) are conditionally independent given \(X\) and, as before, \(T\) and \(A\) are independent given \(X\). Figure 1 summarizes the sampling scheme for all individuals in the study and the available prospective survival information.

We model the dependence of \(S\) on \(X\) using a Cox proportional hazards model, and thus \(S(t|X, \kappa) = \exp\{-\Lambda(t|X, \kappa)\}\) with

\[
d\Lambda(t|X = x, \kappa) = \lambda(t|X = x, \kappa) = \lambda_0(t) \exp(x^T \gamma),
\]
where $\lambda_0$ is the baseline hazard function that depends only on time, and $\kappa = (\lambda_0, \gamma)$.

Defining the indicator variable $R = 1$ for a prevalent case, and $R = 0$ for an incident case, and letting $g(t|X, \kappa) = -dS(t|X, \kappa)/dt$ denote the density corresponding to $S$, the likelihood for the survival data for the $n_1$ incident and $n_2$ prevalent cases is proportional to

$$L_S(\kappa) \propto \prod_{i=n_0+1}^{N} g^R(Y_i|X_i, \kappa) S^{1-\delta_i}(Y_i|X_i, \kappa) |\mu(X_i, \kappa)|^{-R_i}.$$  \hspace{1cm} (8)

Combining the log-likelihood corresponding to (8) for the prospective survival data with the profile log-likelihood $l_p(\alpha, \nu, \beta, \kappa)$ in (6) for the case-control data and the backward time on the prevalent cases yields the full data profile log-likelihood

$$l(\alpha, \nu, \beta, \kappa) = -\sum_{i=1}^{N} \log[1 + \exp(\alpha + x_i^T \beta) + \exp(\nu + x_i^T \beta + \log \mu(x_i, \kappa))]
$$

$$+ \sum_{i=n_0+1}^{n_2+n_1}(\alpha + x_i^T \beta) + \sum_{i=n_0+n_1+1}^{N} \left[ \nu + x_i^T \beta + \log \left( \frac{S(a_i|X_i, \kappa)}{\mu(x_i, \kappa)} \right) \right] + \sum_{i=n_0+1}^{N} \left[ \delta_i \log g(y_i|X_i, \kappa) + (1 - \delta_i) \log S(y_i|X_i, \kappa) \right].$$

(9)

Under a parametric model for $\lambda_0(t)$ in (7) the above profile log-likelihood can be maximized jointly for all parameters using standard optimization. For example, when $\lambda_0$ is a Weibull hazard with shape and scale parameters $\kappa_1$ and $\kappa_2$, respectively, $S(t|X, \kappa) = \exp\left\{- \left(t/\kappa_2\right)^{\kappa_1} \exp(x^T Y) \right\}$ and $\mu(x, \kappa) = \Gamma(k^{-1})/\left(\kappa_1 \kappa_3^{1/k_1} \right) \Gamma^{-1}(k_1) \int_0^{\kappa_2^{k_1}} \exp(-u)u^{(1/k_1)-1}du$ with $\kappa_3 = \kappa_2^{k_1} \exp(x^T Y)$.

However, for a general unspecified baseline hazard function $\lambda_0(t)$ that one wishes to estimate nonparametrically, optimizing (9) becomes computationally extremely difficult, even though all parameters are theoretically identifiable. We therefore develop a two-step generalized method of moment approach for estimation that we discuss next.

## 3 TWO-STEP PARAMETER ESTIMATION

We now propose and study a two-step generalized method of moments approach to estimating $(\alpha, \nu, \beta, \kappa)$ when $S$ is modeled based on the Cox proportional hazard model, that is, $\lambda(t|X, \kappa) = \lambda_0(t) \exp(x^T Y)$, where $\lambda_0$ is an unspecified baseline hazard function. First, we estimate $\kappa = (\lambda_0, \gamma)$ semiparametrically using the prospective survival information from incident and/or prevalent cases. Then we plug $\hat{\kappa}$ into (6) and maximize the pseudo-log-likelihood as a function of the remaining parameters $(\alpha, \nu, \beta)$.

### 3.1 Step 1: Estimate $\kappa = (\lambda_0, \gamma)$ from prospective follow-up data for incident and prevalent cases

We adapt an EM algorithm proposed by Qin et al.\textsuperscript{6} and further modified by Liu et al.\textsuperscript{5} to estimate $\lambda(t|X, \kappa)$ in (7) when follow-up information from prevalent and incident cases is available, to obtain fully efficient estimates of $\kappa$. For comparison we also estimate $\kappa$ based on the standard Cox partial likelihood with left truncation. For ease of exposition we start indexing the cases at index $i = 1$.

**a) Estimating $\kappa$ via an EM algorithm**

The basic idea for the EM algorithm is that for the $i$th prevalent case that is observed, $m_i$ cases were left truncated, that is, are unobserved. The “missing data” for the $i$th prevalent case are thus $O_i^p = \{(T_{il}^+, A_{il}^+), \ldots, (T_{im}^+, A_{im}^+)\}$ where $T_{il}$ and $A_{il}$ are the survival and backward times, respectively, for the $i$th unobserved prevalent case with $T_{il}^+ < A_{il}^+$. The complete data for each prevalent case are $(O, O_i^p)$.

Let $\lambda_j = \lambda_0(t), j = 1, \ldots, k$, where $0 < t_1 < t_2 < \cdots < t_k$ are the observed times of death for all cases (prevalent and incident). The complete data log-likelihood for incident and prevalent cases, based on (7) and (8) is
\[ l_c(\lambda) = \sum_{j=1}^{k} \sum_{i=1}^{n_i+n_j} \left[ I(Y_i = t_j) \left\{ \delta_i (\log \lambda_j + x_i^T \beta_j) - \exp(x_i^T \beta_j) \sum_{p=1}^{j} \lambda_p \right\} + \sum_{i=1}^{n_i} r_i I(T_i^u = t_j) \left\{ \log \lambda_j + x_i^T \beta_j - \exp(x_i^T \beta_j) \sum_{p=1}^{j} \lambda_p \right\} \right] . \] (10)

\( I \) denotes the indicator function that is 1 if the argument is true and 0 otherwise. Conditional on the observed data \( O_i \) for the \( ith \) subject, we write the expectation in the E-step as

\[ \hat{w}_j = E \left[ \sum_{i=1}^{m} I(T_i^u = t_j) | O_i \right] = \frac{\hat{\lambda}_j}{v_i} \left( 1 - \frac{t_j}{\hat{\xi}_j} \right) \omega_{ij}. \] (11)

where \( \omega_{ij} = \lambda_j \exp(x_i^T \beta_j) \exp(-\sum_l^{j\ell} \lambda_l \exp(x_i^T \beta_l)) \), \( v_i = \sum_{i=1}^{k} I(t_i \omega_{ij}) \), and following Qin et al., \( \hat{\xi} = \max(Y_i, \ldots, Y_{n_i+n_j} = t_k) \).

As described in more detail in the Supplemental Material, \( \hat{\xi} \) converges to the true parameter \( \xi \) at a rate that is faster than \( n^{-1/2} \). Therefore we can treat \( \hat{\xi} \) as a fixed and known constant when estimating the remaining parameters.

In the M-step, we maximize the expected complete-data log-likelihood function conditional on the observed data,

\[ Q(\kappa | \kappa^{(u)}) = \sum_{j=1}^{k} \sum_{i=1}^{n_i+n_j} \left[ I(Y_i = t_j) \left\{ \delta_i (\log \lambda_j + x_i^T \beta_j) - \exp(x_i^T \beta_j) \sum_{p=1}^{j} \lambda_p \right\} + w_0^{(u)} r_i \left\{ \log \lambda_j + x_i^T \beta_j - \exp(x_i^T \beta_j) \sum_{p=1}^{j} \lambda_p \right\} \right] . \] (12)

We define vectors of length \( (n_1 + n_2)k \) for the distinct failure times \( T_{(n_1+n_2)k} = (t_1, \ldots, t_k, \ldots, t_1, \ldots, t_k)^T \), the covariates \( X_{(n_1+n_2)k} = (x_1, \ldots, x_1, \ldots, x_{n_1+n_2}, \ldots, x_{n_1+n_2})^T \) and the censoring indicators \( \delta_{(n_1+n_2)k} = (1, \ldots, 1)^T \). Estimates \( \hat{\lambda}_j^{(u)} \) can be computed by fitting a weighted Cox regression model, for example, using the coxph function in R, \( \text{coxph} \) \((\text{Surv}(T_{EM}, \delta_{EM}), \text{weights} = W_{EM}) \) where \( W_{EM} = (1, \ldots, 1, w_{11}, \ldots, w_{1k}, \ldots, w_{(n_1+n_2)1}, \ldots, w_{(n_1+n_2)k})^{T} \) is a vector of length \( (n_1 + n_2) + (n_1 + n_2)k \) with weights estimated in the E-step. Also, \( T_{EM} = (Y_i, \ldots, Y_{n_1+n_2}, T_{(n_1+n_2)k}) \), \( \delta_{EM} = (\delta_i, \ldots, \delta_{n_1+n_2}, \delta_{(n_1+n_2)k}) \) and \( X_{EM} = (x_i, \ldots, x_{n_1+n_2}, X_{(n_1+n_2)k}) \). Estimates \( \hat{\lambda}_j^{(u)} \) in (12) have close-form solutions,

\[ \hat{\lambda}_j^{(u)}(\gamma^{(u)}) = \frac{\sum_{i=1}^{n_1+n_2} \left\{ w_0^{(u)} r_i I(Y_i = t_j) \delta_i \right\}}{\sum_{i=1}^{n_1+n_2} \sum_{i=1}^{k} \left\{ w_0^{(u)} r_i I(Y_i = t_j) \right\} \exp(x_i^T \beta^{(u)})}. \] (13)

**Remark 1.** Our derivations above assume that the backward time \( A \) has a uniform distribution, however, if stationarity of disease incidence in the underlying population is an unreasonable assumption. The EM algorithm can be extended to other parametric distributions for \( A \), as shown in Supplemental Material.

**b) Cox partial likelihood with left truncation**

While estimates \( \hat{\kappa} \) from the EM algorithm are more efficient,\(^5\) one could also estimate \( \kappa \) using the standard Cox partial likelihood.\(^4,7\) This approach does not require making any distributional assumptions for the backward time \( A \) but it yields estimates that are less efficient than those obtained from a full likelihood.\(^8\)

For individual \( i \) the counting process \( N_i(t) \) is defined as \( N_i(t) = I(Y_i \leq t, \delta_i = 1), t \geq 0, i = 1, \ldots, n_1 + n_2 \). Left truncation of the prevalent cases is accommodated in the “at risk process” \( Z_i(t) = I(A_i < t \leq Y_i) \), where \( A_i \) is the backward time if \( i \) is a prevalent case and \( A_i = 0 \) for an incident case. \( Z(t) \) is not monotone decreasing with \( t \) as prevalent cases are at risk only since time \( A \). Letting \( dN_i(t) = N_i(t) - N_i(t-) \) denote the increment of \( N_i \) at time \( t \), the score functions based on the partial likelihood for \( \gamma \) are

\[ U_{11}(\gamma) = \sum_{i=1}^{n_1+n_2} \sum_{t \geq 0} \left\{ x_i \frac{\hat{S}^{(1)}(t; \gamma)}{\hat{S}^{(0)}(t; \gamma)} \right\} dN_i(t) = 0, \] (14)

with \( \hat{S}^{(u)}(t; \gamma) = \sum_{j=1}^{n_1+n_2} Z_j(t) \exp(\gamma^T x_j) x_j^{\delta_j} \), where for a column vector \( a, a^{\otimes 0} = 1, a^{\otimes 1} = a, a^{\otimes 2} = aa^T \). Given \( \hat{\gamma} \), the estimating equations for \( \hat{\lambda}_0(t) \) are
resulting in the Breslow estimate of the cumulative baseline hazard at time \( t \),

\[
\hat{\Lambda}_0(t) = \int_0^t \hat{\lambda}_0(s) \, ds = \sum_{i=1}^{n_0} \sum_{j=1}^{n_i} \frac{\sum_{k=1}^{n_{ij}} \hat{d}_N(t)}{\sum_{k=1}^{n_{ij}} Z_i(t) \exp(\hat{\gamma}_i^T \mathbf{x}_i)},
\]

with 0 < \( t_1 < \cdots < t_k \) denoting the observed event times for all cases.

Finally, after estimating the parameters of the survival distribution via the EM algorithm or based on the Cox partial likelihood with left truncation, for a given covariate \( \mathbf{X} = \mathbf{x} \), we use \( \hat{\kappa} \) and \( \hat{\zeta} = t_k \) in expression (4), and obtain

\[
\mu(\mathbf{X}, \hat{\kappa}) = \int_0^{t_k} \exp[-\hat{\Lambda}_0(t) \exp(\mathbf{x}^T \hat{\gamma})] \, dt = \sum_{j=1}^{k} (t_j - t_{j-1}) \exp[-\hat{\Lambda}_0(t_{j-1}) \exp(\mathbf{x}^T \hat{\gamma})].
\]

### 3.2 Step 2: Estimate \( \theta = (\alpha, \nu, \beta) \) given \( \hat{\kappa} \)

We now treat \( \mu(\mathbf{X}, \hat{\kappa}) \) in (17) as a known function of \( \mathbf{X} \) and estimate the remaining parameters \( (\alpha, \nu, \beta) \) by maximizing the pseudo log-likelihood

\[
l(\alpha, \nu, \beta|\hat{\kappa}) = -\sum_{i=1}^{N} \log[1 + \exp(\alpha + \mathbf{x}_i^T \hat{\beta}) + \exp(\nu + \mathbf{x}_i^T \hat{\beta} + \log \mu(\mathbf{x}_i, \hat{\kappa})] + \sum_{i=\tilde{n}_0+1}^{n_0+n_1} (\alpha + \mathbf{x}_i^T \hat{\beta}) + \sum_{i=\tilde{n}_0+n_1+1}^{n_0+n_2+n_1} (\nu + \mathbf{x}_i^T \hat{\beta}).
\]

**Theorem 1.** Denote the estimator that maximizes (18) by \( \hat{\theta} = (\hat{\alpha}, \hat{\nu}, \hat{\beta})^T \) and the true value by \( \theta_0 = (\alpha_0, \nu_0, \beta_0)^T \). Then \( N^{1/2}(\hat{\theta} - \theta_0) \overset{D}{\sim} N(0, \Sigma \theta^\nu \Sigma^{-1}) \) with \( \Sigma \) and \( \theta^\nu \Sigma^{-1} \) defined in Supplementary Material.

The proof of the theorem is given in Supplementary Material.

Standard deviations and 95% confidence intervals for \( \theta \) can be obtained based on empirical estimates of \( \Sigma \) and \( \theta^\nu \Sigma^{-1} \) or using a bootstrap resampling procedure that samples controls, incident, and prevalent cases with replacement from the respective groups, with fixed sample sizes \( n_0, n_1 \), and \( n_2 \) and then fits steps 1 and 2 of the two-step procedure for each bootstrap sample. Confidence intervals can be computed either based on the bootstrap standard deviations assuming normality, or based on the quantiles of the bootstrap distribution.

### 4 Simulation Study

We assessed small sample bias of our two-step approach and compared its efficiency with several other methods in simulations.

#### 4.1 Data generation

We generated data from a retrospective setting. For controls we obtained \( n_0 \) covariate values from \( \mathbf{X}_0 = (X_{01}, X_{02})^T \sim N(0, \Sigma^0) \), where \( \Sigma_{ii}^0 = 1 \) and \( \Sigma_{ij}^0 = \Sigma_{ji}^0 = 0.5, i \neq j \). For incident cases, we used importance sampling to generate \( n_1 \) covariates from model (2), \( \mathbf{X}_1 = (X_{11}, X_{12})^T \sim f_1 \), as follows. We first generated \( \tilde{n}_1 \) realizations of \( \mathbf{X}_1 \sim N(0, \Sigma^0) \), where \( \tilde{n}_1 >> n_1 \). Then we drew a sample of size \( n_1 \) with replacement where each observation \( \hat{\mathbf{X}}_{1,k}, k = 1, \ldots, n_1 \) was sampled with probability \( \exp(\mathbf{x}_{1,k}^T \beta) / \sum_{j=1}^{n_0+n_1} \exp(\mathbf{x}_{1,j}^T \beta) \) which ensures that the resulting sample arises from distribution \( f_1 \) for \( \beta = (\beta_1, \beta_2)^T = (0, 0)^T \) and \( (1, -1)^T \).

Survival times for incident cases were generated assuming a Weibull baseline hazard function in model (7) with \( \lambda_0(t) = \kappa_1 t^{\kappa_1-1} / \kappa_2^\kappa_1 \), where \( \kappa_1 \) and \( \kappa_2 \) are the shape and scale parameters, respectively.
To obtain covariates for prevalent cases, we first generated $X_2$ from $f_1$ as described above. Given $X_2$, $T$ was drawn from $S$ with a Weibull baseline hazard function $\lambda_0(t)$, and only those samples with $T > A$ were selected, where the backward time $A$ was drawn from a uniform distribution, $A \sim U[0, \xi]$, with $\xi = 30$. This selection procedure tilts the distribution of $X_2$ from $f_1$ to $f_2$ and thus yields $X_2 \sim f_2$. Alternatively, one could use importance sampling with weights $w(x) = \exp[\mathbf{x}^T \beta + \log\{\mu(x; \kappa)\}]$ to generate $X_2$, similar to the incident cases.

The censoring variables for incident and prevalent cases, $C \sim U[0, \tau]$ was generated with different values of $\tau$ to obtain the same amount of censoring among both, incident and prevalent cases. In sensitivity analyses we also let the distribution of $C$ depend on $X$. For incident cases, we set $T = \min(T, C_1)$, and for prevalent cases, $T = A + \min(T - A, C_2)$, that is, for prevalent cases we censored the forward time, which the difference between the total survival time and the backward time. We studied the settings of 10% ($\tau = 5$ for incident and $\tau = 15$ for prevalent cases), 50% ($\tau = 0.6$ for incident and $\tau = 1.5$ for prevalent cases) and 90% censoring ($\tau = 0.05$ for incident and $\tau = 0.15$ for prevalent cases).

The simulation results in all tables are based on 500 replications for each setting.

### 4.2 Analysis methods

We compared the small sample bias and the efficiency of estimates from our two-step approach, implemented using the EM algorithm (“EM” in the tables) or the truncation-adjusted Cox-partial likelihood (“Cox”) in Section 3 to estimates from several different methods. The first one is “joint,” that is, maximizing the full profile likelihood (9) that also incorporates the prospective follow-up data assuming a Weibull baseline hazard jointly for the survival parameters, $(\gamma, \kappa_1, \kappa_2)^T$, and logistic parameters $(\alpha, \nu, \beta)$. We also obtain estimates from maximizing the profile likelihood (6) using only backward time information, termed “IP-CC” for incident/prevalent case-control study where $S$ was parameterized using a Cox model with a Weibull baseline hazard. And lastly, we compute estimates from a standard logistic regression model that simply combines incident and prevalent cases into a single group (“naive”) or uses only incident cases (“IC”).

### 4.3 Results

Table 1 shows estimates (Est) and empirical standard deviations (SDs) and the coverage probabilities (CPs) of 95% Wald-type confidence intervals (CIs) computed using bootstrap standard deviations, as the amount of censoring in the prospective follow-up data of the cases increased from 10% to 90%. The true log-odds ratios were $\beta = (1, -1)$ and the Cox regression log-hazard ratio (HR) parameters were $\gamma = (1, -1)$.

For $n_0 = n_1 = n_2 = 500$ with 10% and 50% censoring, $\hat{\beta}$, and $\hat{\gamma}$ were unbiased for all methods, including all parametric models (joint and IP-CC), since the survival time was generated from an exponential distribution. Not surprisingly, the methods that used prospective follow-up time resulted in much small SDs for the log-HR parameters $\gamma$ ($SD = 0.04$ or $SD = 0.05$) than the IP-CC method ($SD = 0.1$), that only utilizes the backward time of the prevalent cases. The SDs for $\beta$ were virtually the same for all methods ($SD = 0.06$ or $SD = 0.07$; Table 1). The coverage of the 95% CIs was close to nominal for $\hat{\beta}$ for all parameters estimated from methods that accounted for the survival bias in the prevalent cases. 95% CIs based on the naive analysis had 0% coverage for $\hat{\beta}$.

For $n_0 = n_1 = n_2 = 500$ and 90% censoring of the prospective case follow-up times, estimates based on the two-step algorithm with the EM were biased, with $\hat{\gamma} = (0.78, -0.78)$ and $\hat{\beta} = (0.84, -0.85)$, and the coverage of the 95% CIs was less than 50% for all parameters. Estimates $\hat{\gamma}$ from the two-step algorithm with $\hat{\kappa}$ from the Cox partial likelihood were unbiased, with an 8% bias in $\hat{\beta} = (0.92, -0.92)$ (Table 1) and slightly below nominal coverage of the 95% CIs (91.2% for $\hat{\beta}_1$ and 91.8% for $\hat{\beta}_2$).

The small-sample bias for the two-step procedure with the EM algorithm or the Cox partial likelihood decreased as the sample size increased for either prevalent or incident cases (Supplemental Tables 1 and 2, respectively). The bias in $\hat{\beta}$ decreased to about 9% using the EM for $n_0 = n_1 = 500$, $n_2 = 1000$ (Supplemental Table 1), and to 5% for the Cox method and estimates $\hat{\gamma}$ were unbiased for both methods. For $n_0 = 500$, $n_1 = 1000$, $n_2 = 500$ (Supplemental Table 2), the EM-based estimates $\hat{\gamma}$ had a 13% bias while estimates $\hat{\gamma}$ from the Cox partial likelihood were unbiased. The corresponding estimates $\hat{\beta}$ had a 7% and 8% bias for the EM and Cox partial likelihood estimation, respectively. Coverage of the 95% CIs for the EM based estimates when $n_0 = n_1 = 500$, $n_2 = 1000$ was > 90% for all parameters under 50% censoring and it was around 70%
TABLE 1  Estimates (Ests) and empirical standard deviations (SDs) from different estimation methods based on 500 replications under various amounts of censoring for $n_0 = 500$, $n_1 = 500$, $n_2 = 500$. 95% Wald-type CIs were computed using bootstrap SDs from 200 bootstrap samples and then the coverage percentages (CPs) were averaged over 500 replications.

| Method          | $\beta_1 = 1$ | $\beta_2 = -1$ | $\gamma_1 = 1$ | $\gamma_2 = -1$ | $k_1 = 1$ | $k_2 = 1$ |
|-----------------|---------------|----------------|----------------|----------------|------------|------------|
| **10% censoring, $\hat{\xi} = 29.5$** |               |                |                |                |            |            |
| Two-step        | Est (EM)      | 1.00           | -1.01          | 1.03           | -1.03      |            |
|                 | SD (EM)       | 0.07           | 0.07           | 0.04           | 0.04       |            |
|                 | CP (EM)       | 95.0           | 95.8           | 89.6           | 89.0       |            |
|                 | Est (Cox)     | 1.00           | -1.00          | 1.00           | -1.00      |            |
|                 | SD (Cox)      | 0.06           | 0.07           | 0.05           | 0.05       |            |
|                 | CP (Cox)      | 94.2           | 95.6           | 95.0           | 94.60      |            |
| Likelihood      | Est (joint)   | 1.00           | -1.00          | 1.00           | -1.01      | 1.00       | 1.00       |
|                 | SD (joint)    | 0.06           | 0.07           | 0.04           | 0.04       | 0.03       | 0.04       |
|                 | CP (joint)    | 95.0           | 96.0           | 96.0           | 94.8       |            |
| **50% censoring, $\hat{\xi} = 24.4$** |               |                |                |                |            |            |
| Two-step        | Est (EM)      | 1.03           | -1.04          | 0.99           | -0.99      |            |
|                 | SD (EM)       | 0.07           | 0.07           | 0.06           | 0.06       |            |
|                 | CP (EM)       | 93.0           | 93.8           | 94.6           | 95.4       |            |
|                 | Est (Cox)     | 1.00           | -1.00          | 1.01           | -1.00      |            |
|                 | SD (Cox)      | 0.07           | 0.08           | 0.06           | 0.06       |            |
|                 | CP (Cox)      | 95.2           | 95.8           | 95.0           | 95.6       |            |
| Likelihood      | Est (joint)   | 1.00           | -1.00          | 1.01           | -1.01      | 1.01       | 1.01       |
|                 | SD (joint)    | 0.06           | 0.07           | 0.05           | 0.05       | 0.03       | 0.04       |
|                 | CP (joint)    | 95.8           | 96.0           | 95.8           | 95.0       |            |
| **90% censoring, $\hat{\xi} = 23.8$** |               |                |                |                |            |            |
| Two-step        | Est (EM)      | 0.84           | -0.85          | 0.78           | -0.78      |            |
|                 | SD (EM)       | 0.07           | 0.07           | 0.07           | 0.07       |            |
|                 | CP (EM)       | 34.4           | 45.2           | 15.0           | 12.8       |            |
|                 | Est (Cox)     | 0.92           | -0.92          | 1.02           | -1.01      |            |
|                 | SD (Cox)      | 0.12           | 0.12           | 0.15           | 0.13       |            |
|                 | CP (Cox)      | 91.2           | 91.8           | 95.2           | 95.4       |            |
| Likelihood      | Est (joint)   | 1.00           | -1.01          | 1.02           | -1.02      | 1.01       | 1.01       |
|                 | SD (joint)    | 0.07           | 0.07           | 0.06           | 0.06       | 0.04       | 0.06       |
|                 | CP (joint)    | 95.8           | 96.2           | 94.0           | 94.8       |            |
| Likelihood      | Est (IP-CC)   | 1.00           | -1.01          | 1.02           | -1.03      | 1.01       | 1.01       |
|                 | SD (IP-CC)    | 0.07           | 0.07           | 0.10           | 0.10       | 0.09       | 0.13       |
|                 | CP (IP-CC)    | 94.2           | 96.0           | 93.6           | 94.0       |            |
| Logistic        | Est (Naive)   | 0.44           | -0.44          |                |            |            |
|                 | SD (Naive)    | 0.06           | 0.06           |                |            |            |
|                 | CP (Naive)    | 0.0            | 0.0            |                |            |            |
|                 | Est (IC)      | 1.00           | -1.01          |                |            |            |
|                 | SD (IC)       | 0.08           | 0.09           |                |            |            |
|                 | CP (IC)       | 95.0           | 95.2           |                |            |            |

Note: In the population, $(X_1, X_2)^T$ are multivariate normally distributed with mean $(0, 0)$, $\text{Var}(X_1) = \text{Var}(X_2) = 1$, and $\text{Cov}(X_1, X_2) = 0.5$. 
Relative efficiency of $\hat{\beta}_1$ at 10% censoring

Relative efficiency of $\hat{\beta}_2$ at 10% censoring

Relative efficiency of $\hat{\beta}_1$ at 50% censoring

Relative efficiency of $\hat{\beta}_2$ at 50% censoring

Relative efficiency of $\hat{\beta}_1$ at 90% censoring

Relative efficiency of $\hat{\beta}_2$ at 90% censoring

Figure 2 shows the relative efficiency, defined as $RE = \frac{\text{Var}(\hat{\beta})}{\text{Var}(\hat{\beta}_{\text{joint}})}$, the ratio of the variances, for $(\hat{\beta}_1, \hat{\beta}_2)$ estimated using the two-step method with the EM or Cox partial likelihood, and IP-CC methods compared with the joint likelihood method for $n_2 = 250, 500, 750$, and $1000$ prevalent cases, and under 10%, 50%, and 90% censoring for $(\beta_1, \beta_2) = (1, -1)$.

for the log-odds ratio parameters under 90% censoring, and near nominal for $\hat{\gamma}$ (Supplemental Table 1). When $n_0 = 500$, $n_1 = 1000$, $n_2 = 500$, coverage was above 90% for all parameters under 50% and 90% censoring (Supplemental Table 2). For both these sample size settings the methods that used a fully parametric specification of the survival function yielded unbiased estimates of all model parameters.

Supplemental Tables 3 to 5 give results for $\beta = (0, 0)$ for different sample sizes. For 10%, 50%, and 90% censoring, all estimates were unbiased for all choices of sample sizes. For 90% censoring with $n_0 = n_1 = n_2 = 500$, EM-based estimates $\hat{\beta}$ had small approximately 7% bias and $\hat{\gamma}$ had 12% bias. These biases decreased with increasing sample size for both case groups.

Results given in Supplemental Table 6 for a setting with the same sample sizes ($n_0 = 700$, $n_1 = 400$, $n_2 = 200$) and amount censoring (90%) as the real data showed a similar bias in the EM-based estimates as seen for the 90% censoring scenario presented in Table 1.

When the survival time $T$ did not depend on the covariates, that is, $\gamma_1 = \gamma_2 = 0$, all methods, including naively combining the incident and prevalent cases and fitting a logistic model, were unbiased (Supplemental Table 7). However, when we generated data using $\gamma_1 = 0$ and $\beta_2 = 0$ and fit a logistic regression to the naively combined data using only $X_1$ as the covariate, we observed a large bias of 20%.

Figure 2 shows the relative efficiency, defined as $RE = \frac{\text{Var}(\hat{\beta})}{\text{Var}(\hat{\beta}_{\text{joint}})}$, the ratio of the variances, for $(\hat{\beta}_1, \hat{\beta}_2)$ estimated using the two-step method with the EM or Cox partial likelihood, and IP-CC methods compared with the joint likelihood method for $n_2 = 250, 500, 750$, and $1000$ prevalent cases, and under 10%, 50%, and 90% censoring for $(\beta_1, \beta_2) = (1, -1)$.
Robustness studies

To assess the robustness of the methods when the underlying baseline did not have a monotone structure we generated data using two different step-functions for $\lambda_0(t)$ in (7) on the intervals $I_1 = [0, 7]; I_2 = (7, 14); I_3 = (14, 21]; I_4 = (21, 30]$. The first hazard function had values $\lambda_0(t) = 10^{-4}, 10^{-5}, 2 \times 10^{-4}, 0.5 \times 10^{-4}$, and the second one had values $\lambda_0(t) = 10^{-5}, 2.0 \times 10^{-4}, 10^{-5}, 2.0 \times 10^{-4}$, for $t \in I_k, k = 1, \ldots, 4$, respectively. Estimates $\hat{\beta}$ from the EM method were unbiased and comparable to the competing methods. Under both of these baselines, under 90% censoring, $\hat{\gamma}$ from the EM method had a 21% bias. However, $\hat{\beta}$ was unbiased (Supplemental Tables 8 and 9).

We conducted several additional robustness investigations that are presented in the Supplemental Material and Supplemental Tables 10 to 12. These tables and related descriptions in Sections 3 and 4 in the Supplemental Material summarize results for covariate-dependent censoring (Supplemental Table 10), the robustness to violations of the uniform assumption of the backward time $A$ (Supplemental Table 11), sensitivity of the two-step EM algorithm to the estimation of the support of the backward time, $\xi$ (Supplemental Table 12), and further simulations to better understand the performance of the methods when the survival distribution of $T$ does not depend on covariates, or depends on covariates that differ from those in the logistic component of the model. Even when covariates in the logistic model were different from those in the survival model naively combining the cases into a single group resulted in a biased log-odds ratio estimates (Supplemental Table 12). Supplemental Table 13 shows that when the data were generated using a Weibull baseline hazard function, but fit assuming that the baseline hazard function was a constant, that is, an exponential hazard, log-odds ratio estimates of $\beta_1$ for the IP-CC approach had a 16% significant bias for the covariate that also impacted the survival distribution. This highlights some sensitivity of the IP-CC approach to mis-specifications of the parametric baseline hazard function.

5 | DATA EXAMPLE

We analyzed data from a case-control study conducted within the USRTS to assess associations of SNPs in candidate genes with risk of breast cancer. The USRTS, initiated in 1982 by the National Cancer Institute and other institutions, enrolled 146,022 radiologic technologists to study health effects from low-dose occupational radiation exposure. Information on participants’ characteristics, exposures, and prior health outcomes was collected via several surveys conducted between 1984 and 2014, and blood sample collection began in 1999.

The breast cancer case-control study used information from the first two surveys, conducted 1984 to 1989 and 1993 to 1998. Women who answered both surveys and were diagnosed with a breast cancer between the two surveys were considered incident cases and women who answered only one survey and reported a prior breast cancer diagnosis were considered prevalent cases. All cases with blood samples for genetic analysis were included in the study. We analyzed data on 711 controls, 386 incident cases, and 227 prevalent cases, with follow-up information on the cases through December 2008, available through regular linkage with the National Death Index. Only 49 breast cancer cases died during follow-up, corresponding to 92% censoring.

We modeled the survival distribution for the time to death after breast cancer onset using Cox proportional hazards regression (7) with either an unspecified or a Weibull baseline hazard function. After some exploratory analyses, the following covariates were included in the models. For the relative risk component of the survival model: genotype for the SNP rs2981582 (1 if TC/TT, 0 if CC); age at breast cancer diagnosis in five categories ($\leq 22$, $(22, 40]$, $(40, 50]$, $(50, 55]$,
TABLE 2  Estimates and bootstrap standard deviations (SDs) for association of single nucleotide polymorphisms (SNPs) with risk of breast cancer in $n_0 = 711$ controls, $n_1 = 386$ incident, and $n_2 = 227$ prevalent cases sampled from the USRTS cohort.

| Variable                        | Two-step estimation | Likelihood | Standard logistic |
|---------------------------------|---------------------|------------|------------------|
|                                 | EM                  | Cox        | Joint            | IP-CC          | Naive IC       |
| Log-hazard ratios from Cox proportional hazards model |                     |            |                  |                |
| Age at diagnosis                 | 0.66 (0.1)          | 0.76 (0.15)| 0.78 (0.1)       | 0.4 (0.1)      |
| Year first worked                | −0.46 (0.24)        | −0.23 (0.32)| −1 (0.26)       | −1.34 (0.25)   |
| History of heart disease         | 0.6 (0.3)           | 0.74 (0.34)| 0.54 (0.34)      | −0.17 (0.34)   |

Parameters of the Weibull baseline

| $k_1$                             | 4.85 (0.36)          | 1.36 (0.32)       |                  |                |
| $k_2$                             | 45.57 (2.35)         | 11.09 (1.96)      |                  |                |

Log-odds ratios from logistic model

| rs2981582                         | 0.09 (0.11)          | 0.09 (0.11)       | 0.09 (0.11)      | 0.09 (0.11)    | 0.11 (0.13)    |
| rs889312                          | 0.24 (0.11)          | 0.24 (0.11)       | 0.24 (0.11)      | 0.24 (0.11)    | 0.25 (0.11)    | 0.28 (0.13)    |
| rs13281615                         | 0.29 (0.11)          | 0.29 (0.11)       | 0.29 (0.11)      | 0.29 (0.11)    | 0.29 (0.11)    | 0.33 (0.14)    |
| Year first worked                 | 0.08 (0.13)          | 0.08 (0.13)       | −0.12 (0.13)     | 0.06 (0.13)    | −0.39 (0.15)   |
| Family history                    | 0.52 (0.15)          | 0.52 (0.15)       | 0.52 (0.15)      | 0.52 (0.15)    | 0.47 (0.17)    |
| BMI                               | −0.32 (0.1)          | −0.32 (0.1)       | −0.32 (0.1)      | −0.32 (0.1)    | −0.29 (0.11)   |

Note: SDs are estimated based on 500 bootstrap samples in parenthesis.
6 | DISCUSSION

In this article we propose and study a two-step semiparametric method to incorporate prospective follow-up information into estimating log-odds ratios for association with disease incidence for case-control studies that include prevalent cases in addition to or instead of incident cases.

While many authors addressed the issue of length-bias when estimating survival parameters from a prevalent cohort, for example, Zhu et al., the literature on using prevalent cases when samples are ascertained cross-sectionally is limited. Begg and Gray adjusted for survival bias when comparing prevalent cases to controls to estimate incidence odds ratios, using a method of moments approach, but did not use any follow-up information. Maziarz et al. proposed a fully efficient method, the IP-CC approach, to estimate associations of an exposure with disease incidence when a case-control study includes both, incident cases and prevalent cases, but needed to model the backward time for the prevalent cases fully parametrically as only cross-sectional information on the cases was used.

Here, we relax the parametric assumptions and model the survival distribution semiparametrically, using a Cox proportional hazards model. To further improve efficiency of the estimates of the model, we also extended the EM algorithm proposed by Qin et al. and Liu et al. for estimation of a survival distribution to accommodate incident and prevalent cases. Combining the survival distribution estimated with the EM algorithm with the profile-log-likelihood for the two case groups and the controls in a two-step fashion yielded estimates of the log-odds ratio parameters that were as efficient as those from jointly maximizing the profile-log-likelihood and the survival data under a parametrically specified survival distribution for most settings we studied in simulations. However, under 90% censoring or when data were simulated under a strongly decreasing baseline hazard function with comparably large support for the backward time, the EM based estimates were noticeably biased for sample sizes of 500 incident and prevalent cases and 500 controls. These biases tended to disappear with larger sample sizes. Estimation of the log-hazards ratio based on the EM algorithm was typically more efficient than using a Cox partial likelihood, as the EM better utilizes the backward time information.

We have the following explanation for the bias in the model estimates when there is a very high percentage of censoring. As the density of the backward time \( A \) is \( f_A(a) = S(a)/\mu, a > 0, f_A(0) = S(0)/\mu = 1/\mu \). Thus \( \mu \) is the inverse of \( f_A(0+) \) (omitting \( X \) for simplicity). Woodroofe and Sun noticed that in general it is not possible to consistently estimate the value of a nonincreasing density at 0+ nonparametrically and therefore \( \mu \) is not estimable based on the backward time alone. When the proportion of censored prospective observations is very high, most of the information on \( \mu \) comes from the backward times, and therefore \( \hat{\mu} \) and as a consequence estimates of the log-odds ratios exhibit some bias.

Our motivating example, the analysis of incident and prevalent breast cancer cases and controls in the USRTS to assess the impact of SNPs on breast cancer risk, also highlights that when there is a large amount of censoring, adding prospective follow-up information does not improve efficiency of the association estimates much. However, if prospective information is readily available it should be incorporated into the analysis to lessen the reliance on model assumptions that are needed for fitting the IP-CC method with only cross-sectional information.

When the amount of censoring was limited (around 50%), the two-step procedure with the EM algorithm was unbiased and more efficient than estimating the parameters of the survival distribution using the Cox partial likelihood. This gain in efficiency comes from the EM more fully utilizing information on the backward time, \( A \). While we assume that \( A \) had a uniform distribution, which holds true when the disease process is stationary in the population, the EM algorithm and our two-step procedure can be implemented using any parametrically specified distribution for \( A \). Efficiency gains in log-odds ratio estimates gleaned from the survival information could be lessened, however, if many parameters in the distribution of \( A \) need to be estimated. Another practically appealing aspect of the two-step procedure with the EM estimation is that it is very easy to implement using standard survival software and closed-form expressions for the baseline hazard function estimates, and allows incorporating all available information on incident and prevalent cases sampled into a case control study. However, the two-step procedure with the Cox partial likelihood also provided reliable results and was more robust under large amounts of censoring in the data.

In summary, when case-control studies include prevalent cases, using additional follow-up information on cases is recommended to lessen model dependencies and improve efficiency of estimates of association, especially for outcomes where prevalent cases are readily available.

ACKNOWLEDGEMENTS

The authors thank Jerry Reid, Diane Kampa, Allison Iwan, Jeremy Miller, the radiologic technologists who participated in the USRT study, and the reviewers for helpful comments. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).
DATA AVAILABILITY STATEMENT

Data used in this paper are subject to third-party restrictions.

ORCID

Soutrik Mandal https://orcid.org/0000-0002-0625-8999
Ruth M. Pfeiffer https://orcid.org/0000-0001-7791-2698

REFERENCES

1. Schlesselman JJ. Case-Control Studies: Design, Conduct, Analysis. Oxford, UK: Oxford University Press; 1982.
2. Maziarz M, Liu Y, Qin J, Pfeiffer RM. Inference for case-control studies with incident and prevalent cases. Biometrics. 2019;75(3):842-852.
3. Qin J. Inferences for case-control and semiparametric two-sample density ratio models. Biometrika. 1998;85(3):619-630.
4. Cox D. Regression models and life-tables. J Royal Stat Soc Ser B Stat Methodol. 1972;34(2):187-202.
5. Liu H, Ning J, Qin J, Shen Y. Semiparametric maximum likelihood inference for truncated or biased-sampling data. Stat Sin. 2016;26:1087-1115.
6. Qin J, Ning J, Liu H, Shen Y. Maximum likelihood estimations and EM algorithms with length-biased data. J Am Stat Assoc. 2011;106(496):1434-1449.
7. Cox DR. Partial likelihood. Biometrika. 1975;62(2):269-276.
8. Oakes D. The asymptotic information in censored survival data. Biometrika. 1977;64(3):441-448.
9. Breslow N. Discussion of paper by D.R. Cox. J Royal Stat Soc. 1972;34:216-217.
10. Aalen O. Nonparametric inference for a family of counting processes. Ann Stat. 1978;6:701-726.
11. Bhatti P, Doody MM, Alexander BH, et al. Breast cancer risk polymorphisms and interaction with ionizing radiation among US radiologic technologists. Cancer Epidemiol Prev Biomark. 2008;17(8):2007-2011.
12. Zhu H, Ning J, Shen Y, Qin J. Semiparametric density ratio modeling of survival data from a prevalent cohort. Biostatistics. 2017;18(1):62-75.
13. Begg CB, Gray R. Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika. 1984;71(1):11-18.
14. Woodroofe M, Sun JY. A penalized maximum likelihood estimator of \(f(0+)\) when \(f\) is non-increasing. Stat Sin. 1993;3:501-515.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Mandal S, Qin J, Pfeiffer RM. Incorporating survival data into case-control studies with incident and prevalent cases. Statistics in Medicine. 2021;40(28):6295-6308. doi: 10.1002/sim.9183