




# Inference for case-control studies with incident and prevalent cases

Marlena Maziarz <sup>1</sup> | Yukun Liu<sup>2</sup> | Jing Qin <sup>3</sup> | Ruth M. Pfeiffer <sup>1</sup>

<sup>1</sup>National Cancer Institute, National Institutes of Health, Rockville, Maryland

<sup>2</sup>School of Statistics, East China Normal University, Shanghai, China

<sup>3</sup>National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland

## Correspondence

Ruth M. Pfeiffer, National Cancer Institute, National Institutes of Health, Rockville, MD  
Email: pfeiffer@mail.nih.gov

## Abstract

We propose and study a fully efficient method to estimate associations of an exposure with disease incidence when both, incident cases and prevalent cases, i.e., individuals who were diagnosed with the disease at some prior time point and are alive at the time of sampling, are included in a case-control study. We extend the exponential tilting model for the relationship between exposure and case status to accommodate two case groups, and correct for the survival bias in the prevalent cases through a tilting term that depends on the parametric distribution of the backward time, i.e., the time from disease diagnosis to study enrollment. We construct an empirical likelihood that also incorporates the observed backward times for prevalent cases, obtain efficient estimates of odds ratio parameters that relate exposure to disease incidence and propose a likelihood ratio test for model parameters that has a standard chi-squared distribution. We quantify the changes in efficiency of association parameters when incident cases are supplemented with, or replaced by, prevalent cases in simulations. We illustrate our methods by estimating associations of single nucleotide polymorphisms (SNPs) with breast cancer incidence in a sample of controls, incident and prevalent cases from the U.S. Radiologic Technologists Health Study.

## KEYWORDS

density ratio model, empirical likelihood, exponential tilting model, length biased sampling, outcome dependent sampling, survival bias

## 1 | INTRODUCTION

Case-control studies that compare the frequency of exposures in incident cases to that in healthy individuals to assess associations with risk of disease incidence are popular for rare outcomes, as they are more economical than prospective cohorts. However, like all observational studies, case-control studies are also vulnerable to biases that result in distorted estimates of exposures' associations with disease risk. One possible bias, often called *survival bias*, occurs when prevalent cases, i.e., individuals diagnosed with the disease at some prior time point and alive at the time of sampling for the case-control study, are used in addition to, or instead of, individuals newly diagnosed with disease, namely

incident cases. If the exposure also impacts survival after disease onset, the estimated association of an exposure with disease incidence over- or underestimates the true association. This is a particularly serious problem for diseases that are rapidly fatal, as survivors may comprise a very special subgroup of cases. Many epidemiologic textbooks (e.g. Schlesselman, 1982, p. 133) point out that simply including prevalent cases in case-control studies without any adjustment for the survival bias leads to biased estimates of incidence odds ratios.

There are several approaches to correct for survival bias in the analysis of cohorts comprised of prevalent cases that are then followed to an event of interest such as death (e.g. Cheng and Huang, 2014), but only one approach has

been proposed to explicitly correct for survival bias when prevalent cases are compared with controls. Begg and Gray (1987) subtracted a bias term estimated from a survival model for the backward time, the time between disease diagnosis and sampling, from the log odds ratio estimates obtained from a standard logistic model fit to controls and prevalent cases. An approach to allow incorporating information from prevalent cases in addition to incident cases is thus needed to enhance inference based on case-control data for rare diseases like cancer, where prevalent cases become more readily available due to improved treatment.

Our work was motivated by a case-control study conducted within the U.S. Radiologic Technologists Study (USRTS) to assess the associations of single nucleotide polymorphisms (SNPs) with risk of female breast cancer (Bhatti et al., 2008). The USRTS was initiated in 1982 by the National Cancer Institute and other institutions to study health effects from low-dose occupational radiation exposure. Information on participants was collected via several surveys conducted between 1984 and 2014, and blood sample collection began in 1999. As the number of incident breast cancer cases with blood samples for genetic analysis was limited, we developed methods to also include information on prevalent cases, i.e., women whose breast cancers were diagnosed prior to blood sample collection, to estimate odds ratios for the associations of SNPs with breast cancer incidence.

Our work is based on the well known result on the equivalence between the logistic regression model for prospectively collected data and the exponential tilting, or density ratio model, for retrospectively collected data (Qin, 1998). To accommodate data from incident cases, prevalent cases and controls, we discuss a three-sample exponential tilting density ratio model. For prevalent cases, in addition to covariate information, we observe their backward time, i.e., the time between disease diagnosis and sampling. We model the backward time distribution based on a parametric model for the survival time conditional on surviving to time of sampling (Section 2). In Section 3 we derive a semi-parametric likelihood that combines information from controls, incident and prevalent cases. We estimate log odds ratios for the associations between disease incidence and exposures, and parameters in the model for the backward time using empirical likelihood techniques, and derive the asymptotic properties of the estimates. In Section 4, we assess the performance of the method in simulations and study efficiency of the estimates when prevalent cases are used in addition to, or instead of, incident cases in a study under various scenarios. We illustrate the methods with data from the motivating study on the association of breast cancer risk and SNPs among women sampled from the USRTS (Section 5), before closing with a discussion (Section 6).

## 2 | SEMI-PARAMETRIC MODEL FOR CASE-CONTROL STUDIES WITH INCIDENT AND PREVALENT CASES

Let  $D$  denote the disease indicator, with  $D = 1$  for individuals with disease and 0 for those without (controls), and  $X$  denotes a (vector of) covariate(s).

### 2.1 | Background: exponential tilting model

For incident cases and controls, we assume that the association between  $X$  and  $D$  in the population is captured by the prospective logistic model

$$P(D = 1 | X = x) = \frac{\exp(\alpha_0 + x\beta)}{1 + \exp(\alpha_0 + x\beta)}, \quad (1)$$

where  $\alpha_0$  denotes an intercept term, and  $\beta$  the log odds ratio for the association of  $X$  with  $D$ , the parameter of interest. In the general population, the marginal probability of disease is  $\pi = P(D = 1) = \int P(D = 1 | x)f(x)dx$  where  $f(x) = dF(x)/dx$  is the density of  $X$ , that is unspecified.

In a case-control study, independent samples of fixed sizes  $n_0$  and  $n_1$  are drawn from controls ( $D = 0$ ) and cases ( $D = 1$ ), respectively, and then information on the exposure  $X$  is obtained. Due to the retrospective sampling, only the conditional densities  $f_0(x) = f(x | D = 0)$  and  $f_1(x) = f(x | D = 1)$  are observed. Using Bayes' rule, the prospective model in (1), and letting  $\alpha^* = \alpha_0 + \log\{(1 - \pi)/\pi\}$ , we obtain the two-sample exponential tilting (or density ratio) model

$$\begin{aligned} f_1(x) &= \frac{\exp(\alpha_0 + x\beta)}{1 + \exp(\alpha_0 + x\beta)} \frac{f(x)}{\pi} \\ &= f_0(x) \exp(\alpha_0 + x\beta) \left( \frac{1 - \pi}{\pi} \right) = f_0(x) \exp(\alpha^* + x\beta). \end{aligned} \quad (2)$$

Prentice and Pyke (1979) showed that fitting the prospective model (1) to the retrospectively ascertained exposure data yields consistent estimates of  $\beta$  and the corresponding standard errors. Qin (1998) profiled out the baseline distribution  $f_0(x)$  in equation (2) and derived a constrained empirical likelihood to estimate  $\beta$  and the nuisance parameter  $\alpha^*$ . We adapt this profile likelihood method in the next Section to incorporate information on prevalent cases.

### 2.2 | Data and models for prevalent cases

We now assume that in addition to incident cases, on whom exposure information is ascertained at time of diagnosis, we also have information on exposure  $X$  from prevalent cases, i.e., individuals who developed disease previously and are alive at the time of sample selection for the case-control study. To formalize the notion of a prevalent case, let  $T$  denote the (unobserved) survival time from disease diagnosis to death, with a survival function  $S(T | x) = P(T > t | x)$ , and let  $A$

denote the backward time, defined as the time between disease diagnosis and sampling. We only observe prevalent cases who are alive at the time of sampling, i.e., if  $T > A$ . The sampling scheme for controls, incident and prevalent cases is depicted in Figure 1. In what follows we assume that  $S$  belongs to a known parametric family indexed by parameters  $\gamma$  and use the notation  $S(\cdot|x, \gamma)$ .

We now derive the joint distribution of the covariates  $X$  and the observed backward time,  $A$ , among prevalent cases, that we use in the overall case-control likelihood in the next Section, and extend the exponential tilting model in (2). For prevalent cases alive at the time of sampling,

$$f(X=x, A=a | D=1, T>A) = f(X=x | D=1, T>A) f(A=a | X=x, D=1, T>A). \tag{3}$$

Similar to Wang (1991), we assume that disease incidence is stationary over time, and thus follows a Poisson process with a constant rate. Conditional on the total number of events  $n$  observed in an interval  $[0, \xi]$ , the ordered event times  $Z_{(1)}, \dots, Z_{(n)}$  can be treated as order statistics from a uniform distribution,  $U[0, \xi]$  (Theorem 2.3.1, p. 67, Ross (1996)). Thus the backward times  $A_i = \xi - Z_i$  also arise from  $U[0, \xi]$ . Then, assuming that the disease onset times and death times are independent, using equation (2) and Bayes' theorem, the density of  $X$  for prevalent cases is

$$\begin{aligned} f_2(X) &= f(X=x | D=1, T>A) \\ &= \frac{P(T>A | X=x, D=1) f(X=x | D=1)}{P(T>A | D=1)} \\ &= f_1(x) \frac{\int_0^\xi S(a|x, \gamma) da}{\int_X \int_0^\xi S(a|x, \gamma) da f_1(x) dx} \\ &= f_0(x) \exp \{ v^* + x\beta + \log \mu(x, \gamma) \}, \end{aligned} \tag{4}$$

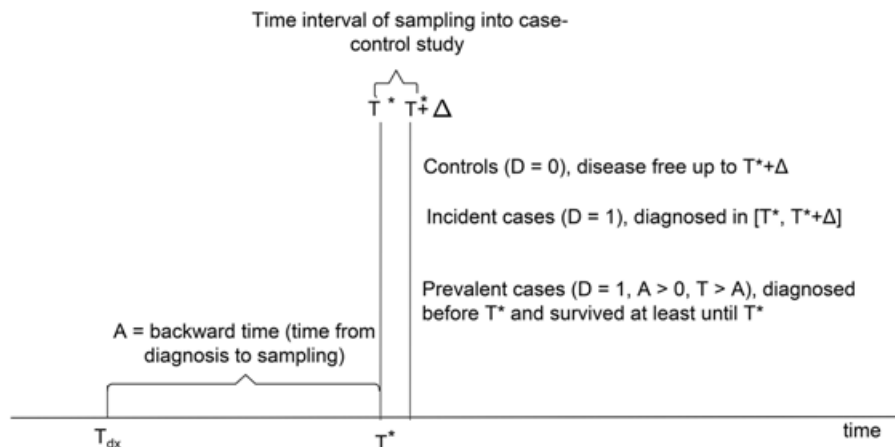
where  $\mu(x, \gamma) = \int_0^\xi S(a|x, \gamma) da$  and  $v^* = \alpha^* - \log \{ \int_X \mu(x, \gamma) f_1(x) dx \}$ . The density of the covariates for the prevalent cases,  $f_2$ , can thus also be expressed in terms of  $f_0(x)$  and a parametric tilting term that, in addition to  $\beta$  and an intercept, depends on the survival distribution  $S$ . Notice that when  $S$  does not depend on  $X$ , the tilting term in (4) depends on  $X$  only through  $X\beta$ , i.e., is the same as for the incident cases, but with a different intercept. However, if  $X$  is related to survival, simply combining prevalent with incident cases and fitting model (1) to the data will lead to biased estimates of  $\beta$ . We use Bayes' theorem and that  $A$  is independent of both,  $X$  and  $T$ , to obtain the conditional density of  $A$  in (3) for  $a \in [0, \xi]$ ,

$$\begin{aligned} f_A(A=a | X=x, D=1, T>A) &= \frac{f(A=a) P(T>a | X=x, D=1)}{P(T>A | X=x, D=1)} = \frac{S(a|x, \gamma)}{\mu(x, \gamma)}. \end{aligned} \tag{5}$$

For  $a \notin [0, \xi]$ ,  $f_A(A=a | X=x, D=1, T>A) = 0$ . In our computations we let  $S(t) = \exp\{-\exp(x\zeta) \int_0^t h_0(s) ds\}$  where  $h_0(t)$  is a constant or Weibull baseline hazard. More flexible parametric models, e.g. splines, could be used as well.

### 3 | SEMI-PARAMETRIC LIKELIHOOD AND INFERENCE

Let  $(x_1, \dots, x_{n_0})'$  denote the covariates for the  $n_0$  controls,  $(x_{n_0+1}, \dots, x_{n_0+n_1})'$  the covariates for the  $n_1$  incident cases and  $(x_{n_0+n_1+1}, \dots, x_N)'$  and  $(a_{n_0+n_1+1}, \dots, a_N)'$  the



**FIGURE 1** Sampling scheme for the IP-case-control study design. For incident cases, disease diagnosis occurs in the case-control sampling period  $[T^*, T^* + \Delta]$ . For prevalent cases, disease is diagnosed at time  $T_{dx}$  before the time  $T^*$  when case-control sampling starts, and information on the backward time  $A$ , i.e., the time between  $T^*$  and  $T_{dx}$ , is also available.

covariates and backward times for the  $n_2$  prevalent cases, where  $N = n_0 + n_1 + n_2$ . Using the models (2) and (4), and the backward time distribution in (5), the likelihood for the controls and the two case groups is

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^N f_0(x_i) \prod_{i=n_0+1}^{n_0+n_1} \exp(\alpha^* + x_i\beta) \\ &\times \prod_{i=n_0+n_1+1}^N \exp\{v^* + x_i\beta + \log \mu(x_i, \gamma)\} \frac{S(a_i | x_i, \gamma)}{\mu(x_i, \gamma)}. \end{aligned} \tag{6}$$

Similar to Qin (1998), we estimate  $p_i = f_0(x_i) = P(X = x_i)$ ,  $i = 1, \dots, N$ , empirically under the following constraints that ensure that the  $f_i$  are, in fact, distributions:  $\sum_{i=1}^N p_i = 1$ ,  $p_i \geq 0$ ;  $\sum_{i=1}^N p_i \exp(\alpha^* + x_i\beta) = 1$ ; and  $\sum_{i=1}^N p_i \exp\{v^* + x_i\beta + \log \mu(x_i, \gamma)\} = 1$ . After maximizing the log-likelihood for  $p_i$  subject to constraints that are accommodated via Lagrange multipliers (see Appendix 1), and letting  $\alpha = \alpha^* + \log(n_1/n_0)$ ,  $v = v^* + \log(n_2/n_0)$ , the profile log-likelihood, referred to as the *IP-case-control likelihood*, for the remaining parameters  $\theta = (\alpha, v, \beta, \gamma)^T$  is

$$\begin{aligned} \ell_p(\theta) &= - \sum_{i=1}^N \log [1 + \exp(\alpha + x_i\beta) + \exp\{v + x_i\beta + \log \mu(x_i, \gamma)\}] \\ &+ \sum_{i=n_0+1}^{n_0+n_1} (\alpha + x_i\beta) \\ &+ \sum_{i=n_0+n_1+1}^N \left[ v + x_i\beta + \log \mu(x_i, \gamma) + \log \left\{ \frac{S(a_i | x_i, \gamma)}{\mu(x_i, \gamma)} \right\} \right]. \end{aligned} \tag{7}$$

**Theorem 1.** Denote the maximum likelihood estimator of  $\theta = (\alpha, v, \beta, \gamma)^T$  in (7) by  $\hat{\theta} = \arg \max_{\theta} \ell_p(\theta)^T$ , and the true value by  $\theta_0 = (\alpha_0, v_0, \beta_0, \gamma_0)^T$ . Under regularity conditions (C1–C3) stated in Supporting Information Section 1, and assuming that the matrix  $\mathbf{V}$  defined in Supporting Information Section 1 is positive definite, as  $N \rightarrow \infty$ ,

- (1)  $N^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, \mathbf{\Omega})$  where  $\mathbf{\Omega} = \mathbf{V}^{-1}\mathbf{\Sigma}\mathbf{V}^{-1}$ , with  $\mathbf{\Sigma}$  defined in Supporting Information Section 4.
- (2) the likelihood ratio  $2\{\sup_{\beta, \gamma} \ell_p(\beta, \gamma) - \ell_p(\beta_0, \gamma_0)\} \xrightarrow{D} \chi_{k_0}^2$ , where  $k_0$  is the dimension of  $(\beta, \gamma)$ ;
- (3) the likelihood ratio for any sub-vector  $\phi$  of  $(\beta, \gamma) \xrightarrow{D} \chi_k^2$ , where  $k$  is the dimension of  $\phi$ .

The proof is given in Supporting Information Section 3. Statement (3) implies that the standard  $\chi^2$  asymptotics also hold for the likelihood ratio test restricted to the parameter  $\beta$ , often of primary interest.

## 4 | SIMULATIONS

We assessed the proposed model and estimation procedure in samples of realistic size, and characterized efficiency of estimates of the log odds ratio parameters  $\beta$  when prevalent cases are used in addition to, or instead of, incident cases in a case-control study.

### 4.1 | Data generation

We generated data directly from the exponential tilting models (2) and (4). We assess the impact of the data generation in Section 4.5. For controls, we simulated  $n_0$  covariates  $\mathbf{X}_0 = (X_{01}, X_{02})^T \sim N(\mathbf{0}, \mathbf{\Sigma}^X)$ , where for  $i = 1, 2$ ,  $\Sigma_{ii}^X = 1$ , and  $\Sigma_{ij}^X = \Sigma_{ji}^X = \rho$ ,  $i \neq j$ , with  $\rho = -0.5, 0$  or  $0.5$ . For incident cases, we generated  $n_1$  values  $\mathbf{X}_1 \sim N(\mathbf{\Sigma}^X \beta, \mathbf{\Sigma}^X)$ , where  $\beta^T = (0, 0), (1, 1), (1, -1)$  or  $(-1, -1)$ . To simulate  $\mathbf{X}$  for  $n_2$  prevalent cases, we first generated  $\tilde{n}_2 \gg n_2$  values  $\tilde{\mathbf{X}}_k \sim N(\mathbf{0}, \mathbf{\Sigma}^X)$ . For each  $\tilde{\mathbf{x}}_k$  we computed a weight  $\tilde{w}_2(\tilde{\mathbf{x}}_k) = \exp\{\tilde{\mathbf{x}}_k^T \beta + \log \mu(\tilde{\mathbf{x}}_k, \gamma)\}$ , and then drew a sample of size  $n_2$  with replacement, where each  $\tilde{\mathbf{x}}_k$  was sampled with probability  $\tilde{w}_2(\tilde{\mathbf{x}}_k) / \sum_j \tilde{w}_2(\tilde{\mathbf{x}}_j)$ . The resulting sample has density  $f_2$  as in equation (4).

The survival distribution was  $S(t | x, \gamma) = \exp\{-(t/\kappa_2)^{\kappa_1} \exp(\mathbf{x}^T \zeta)\}$ , where  $\kappa_1$  and  $\kappa_2$  are the shape and scale parameters of a Weibull baseline hazard  $h_0(t) = (\kappa_1/\kappa_2)(t/\kappa_2)^{(\kappa_1-1)}$ . The parameters associated with  $\mathbf{X}$  were  $\zeta^T = (\zeta_1, \zeta_2) = (0, 0), (1, 1), (1, -1)$  or  $(-1, -1)$ . Then,  $\mu(x, \gamma) = \Gamma(1/\kappa_1) / (\kappa_1 \psi^{(1/\kappa_1)}) \left\{ \Gamma^{-1}(1/\kappa_1) \int_0^{\psi \xi^{\kappa_1}} \exp(-u) u^{(1/\kappa_1-1)} du \right\}$  where  $\psi = \kappa_2^{-\kappa_1} \exp(\mathbf{x}^T \zeta)$ . The expression in the curly brackets is a cumulative Gamma distribution function with shape parameter  $\kappa_1^{-1}$  and scale parameter one, which can be evaluated using standard statistical software. To generate backward times for the prevalent cases, we let  $\kappa_1 = 1$ , generated  $U_i \sim U(0, 1)$  and computed  $A_i = (1/\psi)[- \log\{1 - U_i \psi \mu(x_i, \gamma)\}]$ ,  $i = 1, \dots, n_2$ .

Parameter estimates (Est), empirical standard deviations ( $\text{SD}_{\text{emp}}$ ) and standard deviations ( $\text{SD}_{\text{asy}} = \hat{\mathbf{\Omega}}^{1/2}$ ) were based on  $K = 1000$  replications for each setting. We estimated  $\mathbf{\Omega}$  in Theorem 1 by  $\hat{\mathbf{\Omega}} = \sum_{i=1}^K (\hat{\mathbf{V}}^{-1} \hat{\mathbf{\Sigma}} \hat{\mathbf{V}}^{-1}) / K$ , where  $\hat{\mathbf{V}}$  is the numerical estimate of the Hessian, and  $\hat{\mathbf{\Sigma}}$  the sum of the empirical covariance matrices of the scores for controls, incident and prevalent cases.

### 4.2 | Adding an increasing number of prevalent cases

We first examined the efficiency of parameter estimates when  $n_2 = 500$  or  $1000$  prevalent cases were added to a study with  $n_0 = 500$  controls and  $n_1 = 500$  incident cases.

For  $\beta = (0, 0)$  and  $(1, -1)$ , both with  $\zeta = (1, -1)$ ,  $\rho = 0.5$  and  $\xi = 25$ , all parameter estimates were unbiased, and

the asymptotic and empirical standard deviation estimates agreed well (Table 1). Efficiency results in terms of the ratio of the variance of  $\hat{\beta}$  estimated using the original 500 incident cases and controls, compared to the variance of  $\hat{\beta}$  when  $n_2$  prevalent cases were added, for all combinations of  $\beta$ ,  $\zeta$  and  $\rho$  are shown in Figure 2. As the number  $n_2$  of prevalent cases increased from 0 to 1000, efficiency gains for  $\beta = (0, 0)$  were modest and did not depend on the values of  $\zeta$  or  $\rho$  (Figure 2a); the standard deviations ( $SD_{emp}$ ) for  $\hat{\beta}$ 's decreased slightly as  $n_2$  increased (Table 1a). Efficiency gains were somewhat more noticeable for  $\beta = (1, -1)$  (Table 1b). E.g., for  $\beta = (1, -1)$ ,  $\zeta = (1, -1)$ , and  $\rho = -0.5$ , the ratio of the variance of  $\hat{\beta}$  based on 500 incident cases and controls alone was almost three times larger than the variance after adding 1000 prevalent cases (Figure 2b, Supporting Information Table S2). Additional results are given in Supporting Information Tables S1–S6.

### 4.3 | Increasing the proportion of prevalent cases

Figure 3 shows the efficiency of estimates  $\hat{\beta}$  when the total number of cases was fixed at  $n_1 + n_2 = 500$ , but the proportion of prevalent cases increased,  $n_2/(n_1 + n_2) = 0, 0.2, 0.5, 0.8$

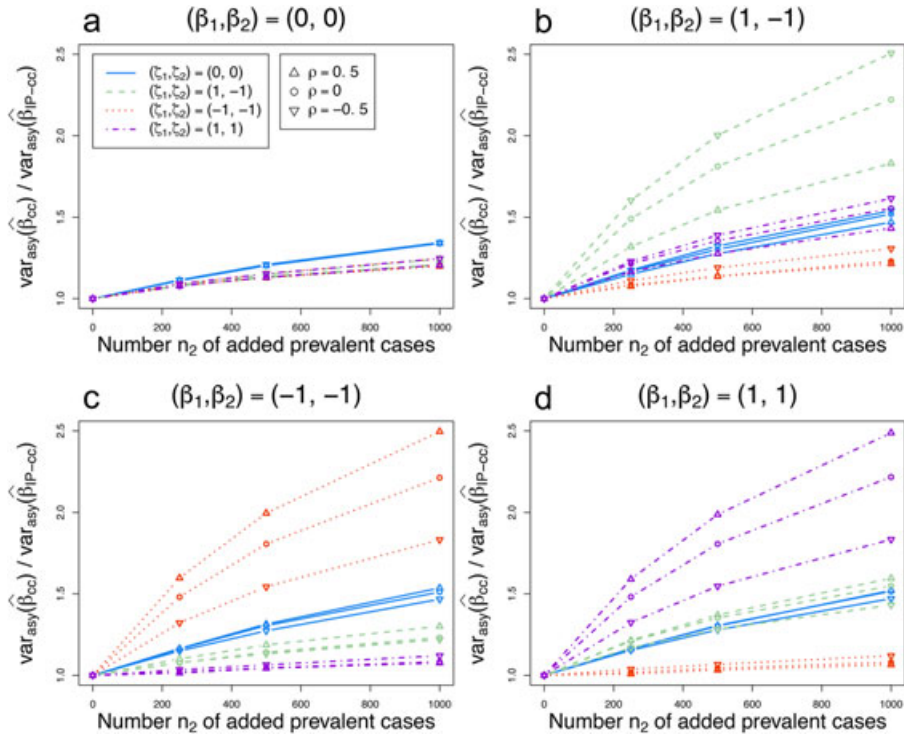
or 1.0. The number of controls was  $n_0 = 500$ . Replacing incident with prevalent cases generally led to an appreciable loss of efficiency of  $\hat{\beta}$ . This was especially apparent when  $\beta = 0$ , where all the ratios  $\text{var}(\hat{\beta}_{cc})/\text{var}(\hat{\beta}_{IP-cc})$  were below one. However, similar to the results in Section 4.2, for some parameter settings there was some gain in efficiency of  $\hat{\beta}$  when prevalent, instead of incident, cases were used. E.g., for  $\beta = (1, -1)$  and  $\zeta = (1, -1)$  with  $\rho = -0.5$ , the  $SD_{emp}$  for  $\hat{\beta}$ 's decreased from 0.107 to 0.091 as the proportion of prevalent cases increased from 0 to 100% (Supporting Information Table S9, Figure 3, and Supporting Information Tables S7–S13).

### 4.4 | Efficiency of $\hat{\beta}$ for added prevalent versus incident cases

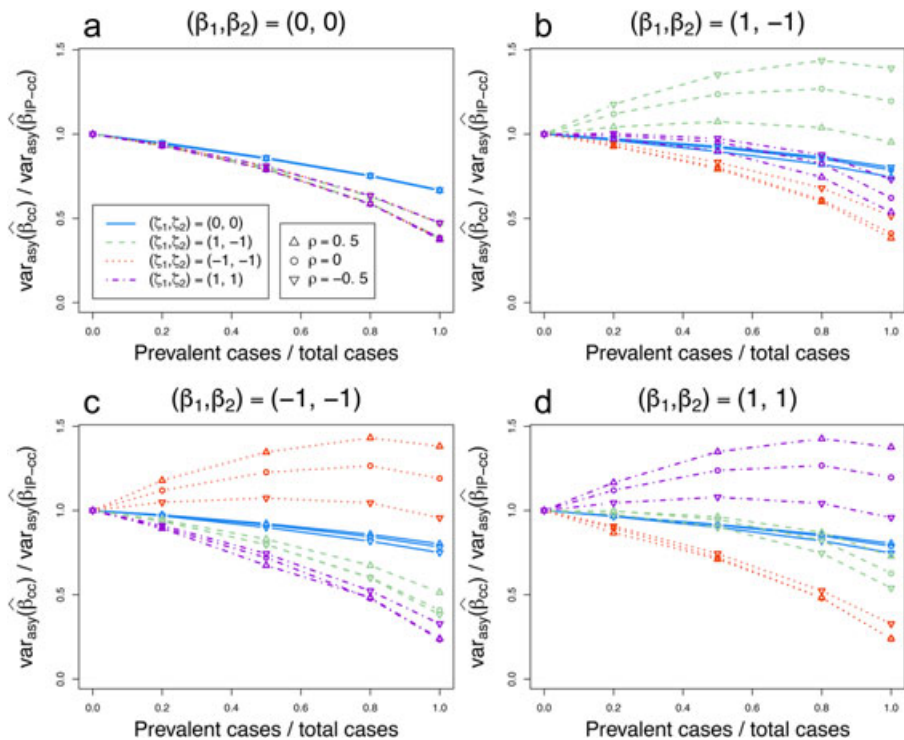
When designing a study, an investigator may have the choice of including additional incident or additional prevalent cases, possibly associated with different costs. We thus investigated differences in efficiency of  $\hat{\beta}$  when adding either incident or prevalent cases to a study comprised of a “base sample” of 500 controls and 500 incident cases. We first added from 20 to 1000 incident cases in increments of 20 to

**TABLE 1** Estimation of the log odds ratios ( $\beta_1, \beta_2$ ) and survival parameters ( $\kappa_1, \kappa_2, \zeta_1, \zeta_2$ ) when the sample size of the prevalent cases varies:  $n_2 = 0, 500, 1000$ , with  $n_0 = n_1 = 500$ . The true log odds ratios were  $\beta = (0, 0)$  for (a) and  $\beta = (1, -1)$  for (b). For both (a) and (b), data were generated with  $\rho = 0.5$  and  $\xi = 25$ . Estimates (Est), empirical standard deviations ( $SD_{emp}$ ), and standard deviation estimates based on the asymptotic covariance matrix ( $SD_{asy}$ ) are based on 1000 replications of the simulation.

(a)	$\alpha^*$	$v^*$	$\beta_1 = 0$	$\beta_2 = 0$	$\kappa_1 = 1$	$\kappa_2 = 1$	$\zeta_1 = 1$	$\zeta_2 = -1$
$n_0 = 500, n_1 = 500, n_2 = 0$								
Est	0.000		0.001	-0.002				
$SD_{asy}$	0.004		0.073	0.074				
$SD_{emp}$	0.003		0.073	0.073				
$n_0 = 500, n_1 = 500, n_2 = 500$								
Est	0.001	-0.473	0.002	0.001	1.017	1.011	1.021	-1.019
$SD_{asy}$	0.004	0.105	0.068	0.068	0.093	0.141	0.100	0.100
$SD_{emp}$	0.002	0.101	0.066	0.068	0.094	0.138	0.105	0.103
$n_0 = 500, n_1 = 500, n_2 = 1000$								
Est	0.001	0.223	0.000	-0.000	1.007	1.003	1.007	-1.008
$SD_{asy}$	0.003	0.077	0.066	0.066	0.065	0.101	0.071	0.071
$SD_{emp}$	0.002	0.076	0.069	0.065	0.065	0.100	0.070	0.072
(b)	$\alpha^*$	$v^*$	$\beta_1 = 1$	$\beta_2 = -1$	$\kappa_1 = 1$	$\kappa_2 = 1$	$\zeta_1 = 1$	$\zeta_2 = -1$
$n_0 = 500, n_1 = 500, n_2 = 0$								
Est	-0.504		1.006	-1.010				
$SD_{asy}$	0.051		0.089	0.089				
$SD_{emp}$	0.049		0.089	0.087				
$n_0 = 500, n_1 = 500, n_2 = 500$								
Est	-0.501	-0.000	1.004	-1.001	1.017	1.014	1.020	-1.019
$SD_{asy}$	0.038	0.095	0.071	0.071	0.087	0.127	0.097	0.097
$SD_{emp}$	0.037	0.094	0.070	0.073	0.090	0.127	0.100	0.099
$n_0 = 500, n_1 = 500, n_2 = 1000$								
Est	-0.500	0.697	1.002	-1.002	1.007	1.004	1.008	-1.007
$SD_{asy}$	0.032	0.067	0.066	0.066	0.060	0.090	0.068	0.068
$SD_{emp}$	0.032	0.066	0.067	0.065	0.061	0.089	0.070	0.071



**FIGURE 2** Efficiency of  $\hat{\beta}_{\text{IP-CC}}$  when  $n_2 = 250, 500$  or  $1000$  prevalent cases (shown on the x-axis) are added to  $n_0 = 500$  controls and  $n_1 = 500$  incident cases (denoted by  $\hat{\beta}_{\text{IP-CC}}$ ), compared to those estimated from controls and incident cases only (denoted by  $\hat{\beta}_{\text{CC}}$ ). Asymptotic variances are used for both estimates. The ratios of the variance estimates are based on 1000 replications of the simulation.



**FIGURE 3** Efficiency of  $\hat{\beta}_{\text{IP-CC}}$  compared to an estimate based on controls and incident cases only ( $\hat{\beta}_{\text{CC}}$ ) as the proportion of prevalent cases out of the total number of cases,  $n_2 / (n_1 + n_2)$  (on the x-axis) increases, for fixed  $n_1 + n_2 = 500$ , and  $n_0 = 500$  controls. Asymptotic variances are used for both estimates. The ratios of the variance estimates are based on 1000 replications of the simulation.

the base sample, and estimated  $\text{var}_{\text{asy}}(\hat{\beta}_{cc})$ . Then, for each value of  $\text{var}_{\text{asy}}(\hat{\beta}_{cc})$ , we found the number  $n_2$  of prevalent cases (increasing in increments of 20) that, if added to the  $n_0 = 500$  controls and  $n_1 = 500$  incident cases, resulted in  $\text{var}_{\text{asy}}(\hat{\beta}_{IP-cc})/\text{var}_{\text{asy}}(\hat{\beta}_{cc}) \approx 1$ .

For most settings, using prevalent cases led to less efficient estimates of  $\beta$ , indicated by lines above the 45° (gray dot-dashed) line in Figure 4, that corresponds to equal variance for the same number of added incident or prevalent cases. This loss of efficiency was particularly apparent when  $\beta = (0, 0)$ , where even for  $\zeta = (0, 0)$ , approximately  $n_2 = 400$  prevalent cases yielded the same variance of  $\hat{\beta}$  as 200 additional incident cases. For some settings a prevalent case provided more information than an additional incident case, as indicated by the lines below the 45° line. For example, for  $\beta = (-1, -1)$  and  $\zeta = (-1, -1)$ , using  $n_2 = 200$  prevalent cases resulted in the same variance of  $\hat{\beta}$  as adding 400 incident cases to the base study sample.

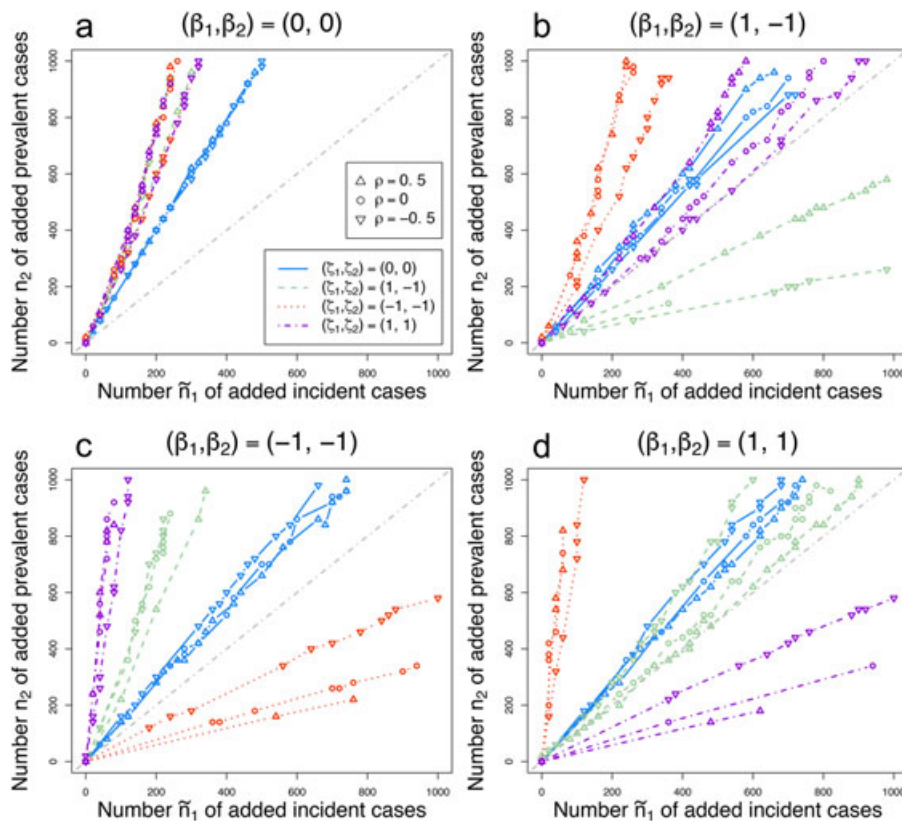
### 4.5 | Robustness studies

We assessed the impact of various violations of assumptions and different sampling schemes on our method.

*Misspecification of S:* As our method requires specifying a parametric survival distribution  $S$  to model the backward time, we examined its robustness to misspecification of  $S$ .

First, we generated backward times based on  $S$  that had a proportional hazards form with a Weibull baseline and was a function of three covariates  $X_1, X_2, X_3$ , but we omitted  $X_3$  in fitting the model. When  $X_3$  was uncorrelated with  $X_1$  and  $X_2$ ,  $\hat{\beta}$  and  $\hat{\zeta}$  were unbiased and there was no loss of efficiency (data not shown). When  $X_3 = 0.5X_1 + 0.5X_2 + \epsilon, \epsilon \sim N(0, 0.25)$ , parameter estimates of  $S$  were biased, but  $\hat{\beta}$  was not, and we saw no loss of efficiency. E.g., when  $\beta = (1, -1)$  and  $n_2 = 1000$  prevalent cases were added to 500 incident cases and 500 controls,  $\hat{\beta} = (1.003, -1.002)$ , but  $\hat{\zeta} = (0.373, -1.117)$  instead of  $\zeta = (1, -1)$  (Supporting Information Table S14).

We also assessed the robustness to misspecification of the baseline hazard of  $S$ . We simulated the backward time for prevalent cases with a Weibull baseline hazard with shape  $\kappa_1 = 3$  and scale  $\kappa_2 = 25$  or a piecewise-constant baseline hazard (details on the simulations are given in Supporting Information Section 8.1.3), and fit the IP-case-control likelihood (7) using either a Weibull or piecewise-constant baseline hazard. We generated the data with  $\beta = (0, 0)$  or  $(1, -1)$  and  $\zeta = (0, 0)$  or  $(1, -1)$  with sample sizes  $n_0 = n_1 = n_2 =$



**FIGURE 4** Number  $n_2$  of prevalent cases on the y-axis, that when added to  $n_0 = 500$  controls and  $n_1 = 500$  incident cases yields the same efficiency of  $\hat{\beta}_{IP-cc}$  as  $\hat{\beta}_{cc}$  when  $\tilde{n}_1$  additional incident cases (on the x-axis) are added to  $n_0 = 500$  controls and  $n_1 = 500$  incident cases. Efficiency  $= \text{var}_{\text{asy}}(\hat{\beta}_{IP-cc})/\text{var}_{\text{asy}}(\hat{\beta}_{cc})$ . The variance estimates are based on 500 replications of the simulation.

500. Estimates of  $\beta$  were unbiased, as were estimates of  $\zeta$  (Supporting Information Table S15).

*Violation of the uniform assumption for the backward time distribution:* To assess the impact of the violation of the assumption that a homogeneous Poisson process gives rise to the disease onset times, we generated  $A$  from a truncated exponential distribution on  $[0, \xi = 25]$ , with density  $f(a) = \lambda \exp(-a\lambda) / \{1 - \exp(-\xi\lambda)\}$ , but fit the model assuming that  $A$  arose from a uniform distribution. The truncated exponential distribution reduces to the uniform distribution when  $\lambda = 0$ . When  $\lambda$  was small,  $\lambda = 0.05$  or  $0.1$ , there was a less than 5% bias in estimates of  $\beta$  (Supporting Information Table S16). For  $\lambda = 0.5$ , corresponding to  $E(A) = 1.98$  instead of  $E(A) = 12.3$  for a uniform  $A$ , the bias was up to 15%.

*Nested case-control sampling:* Here we generated the data from a prospective cohort. We first drew covariates  $\mathbf{X}_i = (X_{i1}, X_{i2}) \sim N(\mathbf{0}, \Sigma^X)$  for  $i = 1, \dots, 800,000$ , cohort members and then generated ages of disease onset from an exponential model where parameters were chosen to yield a 2% disease incidence rate in the age interval  $[0, 70]$ . We allowed for competing mortality and independent censoring (details in Supporting Information Section 8.1.3). We sampled  $n_1$  incident cases and  $n_0 = n_1$  controls, individually matched to cases on age, and in some settings, on an additional binary covariate. The prevalent cases were sampled from incident cases not selected into the case-control study, for whom a randomly generated uniform backward time  $A$  was less than a randomly generated survival time  $T$ . Estimated log odds ratio were unbiased for the hazard ratio parameters, both for  $\beta = (0, 0)$  and  $(1, 1)$  (Supporting Information Table S17).

## 5 | DATA EXAMPLE

We now analyze data from a case-control study conducted within the USRTS to assess associations of SNPs in candidate genes with breast cancer risk (Bhatti et al., 2008). This study used information from the first two surveys, conducted between 1984–1989, 1993–1998. Incident cases were women who answered both surveys and were diagnosed with a primary breast cancer between the two surveys. Controls were frequency matched to incident cases by 5-year categories of year of birth. Prevalent cases were women who answered only one of the surveys and who reported a prior breast cancer diagnosis. Their backward time was defined as the difference between the year of the survey and year of diagnosis. All cancer diagnoses were confirmed using pathology or medical records.

The covariates used in our analysis were: age at diagnosis (cases) or selection (controls) (in categories:  $\leq 22$ , (22, 40], (45, 50], (50, 55],  $>55$ ); the year when the woman started working as a radiation technologist (1 if  $\leq 1955$ , 0 if  $>$

1955); smoking status (1 if former/current, 0 if never); history of breast cancer among first degree relatives (yes/no); BMI ( $\text{kg}/\text{m}^2$ ) during their 20s (in categories:  $\leq 20$ , (20–25],  $>25$ ); age–BMI interaction (coded as BMI during 20s for women diagnosed at  $\geq 50$ , and 0 otherwise, to capture the age dependent effect of BMI on breast cancer risk); history of heart disease (yes/no); alcohol consumption (1 if  $\geq 7$  drinks/week, 0 otherwise); and genotype for three SNPs: rs2981582 (1 if TC/TT, 0 if CC); rs889312 (1 if CA/CC, 0 if AA); and rs13281615 (1 if GG/GA, 0 if AA). We used data from 663 controls, 345 incident cases, and 213 prevalent cases with complete covariate information, however, multiple imputation could have been used to handle missing data. The prevalent cases were older than incident cases and controls, more likely to have started work as a radiation technologist before 1955, to be current smokers, and to have a first-degree relative with breast cancer (Supporting Information Table S18).

We compared log odds ratio estimates ( $\hat{\beta}$ s) from the following models: (A) *Incident model*: standard logistic regression model fit to incident cases only and controls; (B) *Naïve model*: standard logistic regression fit to controls and incident plus prevalent cases combined without accounting for survival bias in the prevalent cases; (C) *IP-case-control*: IP-case-control likelihood (7) fit to incident cases, controls, and prevalent cases accounting for survival bias.

The covariates in the logistic models were the 3 SNPs, age at diagnosis or selection (fit with a trend), year first worked, family history, BMI in 20s (fit with a trend), BMI in 20s (50+), and alcohol consumption. The survival sub-model in (C) was a Weibull model, with the same covariates as in the logistic sub-model plus smoking status and history of heart disease. The support of the backward time was 0 to  $\xi = 40$ , where  $\xi$  was chosen to be larger than the largest observed backward time (35 years). We used jackknife standard errors (SEs) to compute 95% confidence intervals (CIs), assuming normality of  $\hat{\beta}$  or log hazard ratio ( $\log(\text{HR})$ ) estimates.

For model (A), the covariates significantly associated with breast cancer incidence were (95% CIs in parentheses; Table 2): SNP rs13281615,  $\hat{\beta} = 0.40(0.12, 0.69)$ , year first worked,  $\hat{\beta} = -0.84(-1.23, -0.45)$ , family history of breast cancer,  $\hat{\beta} = 0.54(0.25, 0.83)$ , and BMI in ones 20s,  $\hat{\beta} = -0.34(-0.60, -0.08)$ . For model (B), the significant covariates were (Table 2): SNP rs13281615,  $\hat{\beta} = 0.34(0.10, 0.58)$ , family history of breast cancer,  $\hat{\beta} = 0.57(0.32, 0.83)$ , and BMI in ones 20s,  $\hat{\beta} = -0.37(-0.59, -0.14)$ . Model B estimates were attenuated compared to those from model A.

For model (C), the covariates associated with breast cancer incidence were: SNP rs13281615,  $\hat{\beta} = 0.32(0.05, 0.58)$ , year first worked,  $\hat{\beta} = -0.34(-0.65, -0.03)$ , family history of breast cancer,  $\hat{\beta} = 0.53(0.27, 0.79)$ , and BMI in ones 20s,  $\hat{\beta} = -0.34(-0.57, -0.11)$ . The  $\beta$  estimates in the IP-case-control model for rs981782, rs889312 and BMI in 20s were close to



**TABLE 2** Estimated log odds ratios ( $\hat{\beta}$ s) and log hazard ratios (HRs) and jackknife standard errors (SEs) for the association of rs2981582, rs889312, and rs13281615 adjusted for other potential risk factors from three models: *Incident model*: incident cases and controls, estimates from a logistic model; *Naïve model*: incident and prevalent cases combined and controls, estimates from a logistic model; *IP-case-control model*: incident cases, prevalent cases and controls, estimates accounting for survival bias based on the likelihood equation (7).  $\kappa_1$  and  $\kappa_2$  are Weibull baseline hazard shape and scale parameters, respectively.

	<b>Incident (A)</b> $n_I/n_C = 345/663$ $\beta$ (SE)	<b>Naïve (B)</b> $n_{I+P}/n_C = 558/663$ $\beta$ (SE)	<b>IP-case-control (C)</b> $n_I/n_P/n_C = 345/213/663$ $\beta$ (SE)
rs2981582	0.17 (0.14)	0.12 (0.12)	0.14 (0.13)
rs889312	0.23 (0.14)	0.19 (0.12)	0.24 (0.13)
rs13281615	0.40 (0.15)	0.34 (0.12)	0.32 (0.13)
Age at diagnosis/selection	0.13 (0.07)	-0.06 (0.06)	0.06 (0.06)
Year first worked	-0.84 (0.20)	0.02 (0.15)	-0.34 (0.16)
Family history	0.54 (0.15)	0.57 (0.13)	0.53 (0.13)
BMI in 20s	-0.34 (0.13)	-0.37 (0.11)	-0.34 (0.12)
Age-BMI interaction	0.23 (0.21)	0.22 (0.18)	0.16 (0.20)
7+ alcoholic drinks/week	0.13 (0.20)	0.11 (0.17)	-0.004 (0.20)
			log(HR) (SE)
rs2981582			0.05 (0.23)
rs889312			0.21 (0.19)
rs13281615			-0.12 (0.22)
Age at diagnosis/selection			0.48 (0.10)
Year first worked			-1.48 (0.28)
Ever smoker			-0.14 (0.17)
Family history			-0.19 (0.15)
BMI in 20s			0.09 (0.16)
Age-BMI interaction			-0.24 (0.31)
History of heart disease			0.02 (0.41)
7+ alcoholic drinks/week			-0.42 (0.35)
$\kappa_1, \kappa_2$ , Est (SE)			1.58 (0.32), 11.15 (2.13)

those estimated from the incident model, with smaller standard errors. The  $\hat{\beta}$  for age at diagnosis and year first worked were somewhat lower than the estimates of model (A) (Table 2). Age at diagnosis and year first worked were significantly associated with the backward time, with  $\log(\text{HR}) = 0.48$  (0.28, 0.68) and  $\log(\text{HR}) = -1.48$  (-2.03, -0.93), respectively. Not surprisingly, the baseline hazard increased with increasing backward time.

Based on a likelihood ratio test with an asymptotic  $\chi^2_6$  cut-off value, the IP-case control model with the three SNPs in the logistic and the survival models fit the data significantly better than a model without the SNPs ( $p = 0.033$ ).

## 6 | DISCUSSION

The distribution of exposures among prevalent cases, i.e., individuals with a prior disease diagnosis who are alive at the time of sampling for a case-control study, typically differs from that among incident cases. Thus naively combining prevalent and incident cases in the analysis of case-control data leads to biased estimates of log odds ratios for association. We propose a semi-parametric model to incorporate covariates and the observed backward time from prevalent

cases, to obtain unbiased estimates of exposure-disease association. We extend the exponential tilting model to accommodate two case groups and one control group, that we assume is an appropriate comparison group for the incident cases. We provide a semi-parametric method for estimation based on empirical likelihood (Qin and Lawless, 1994; Qin, 1998).

Many authors dealt with the issue of length-bias when estimating survival parameters based on a prevalent cohort (e.g. Cook and Bergeron, 2011; Huang and Qin, 2012; Zhu, 2017). However, few publications use prevalent cases when samples are ascertained cross-sectionally. Without using any information on follow-up, Chan (2013) estimated the impact of a covariate on the survival distribution in a log-linear model by showing that the covariate sampling distribution of prevalent cases compared to incident cases could be expressed using an exponential tilting model. To our knowledge only Begg and Gray (1987) addressed adjusting for survival bias when comparing prevalent cases to controls to estimate incidence odds ratios, again, not using any follow-up information. They modeled the backward time distribution based on an accelerated failure time model for survival and estimated the parameters using quasi-likelihood techniques. Incidence log odds ratios were then estimated by subtracting a bias term

from the log odds ratios from a standard logistic model fit to controls and prevalent cases.

In contrast to the approach by Begg and Gray (1987), our semi-parametric likelihood yields root  $N$  consistent and fully efficient estimates of the incident log odds ratios. We show that the corresponding likelihood ratio statistic has a standard asymptotic  $\chi^2$  distribution, which makes the test practically useful. Based on simulations, the efficiency gains or losses when prevalent cases are added to, or used instead of, incident cases depend on the ratio of the incident to prevalent cases, and the correlation among the covariates in the incidence and survival sub-models. Surprisingly, in some settings, prevalent cases were more informative than incident cases, which warrants further investigation.

A limitation of our approach is that the model for the backward time is fully parametric. However, based on simulations, the estimates of the log odds ratios were not affected by reasonable misspecification of this distribution. Our method is thus very appealing in settings where recall bias for the main exposure is unlikely and the number of available incident cases is limited.

## ACKNOWLEDGEMENTS

M. Maziarz and Y. Liu contributed equally to this work. We thank Michele Doody, National Cancer Institute, for providing the data, Jerry Reid, American Registry of Radiologic Technologists, the radiologic technologists who participated, and the Co-Editor, the AE, and two reviewers for helpful suggestions. Dr Liu was supported by NNSFC (11771144, 11501354, and 11501208) and the 111 project (B14019). This work used the NIH HPC Biowulf cluster (<http://hpc.nih.gov>).

## ORCID

Marlena Maziarz  <http://orcid.org/0000-0003-1417-6050>  
 Jing Qin  <http://orcid.org/0000-0003-2817-6326>  
 Ruth M. Pfeiffer  <http://orcid.org/0000-0001-7791-2698>

## REFERENCES

- Begg, C. B. and Gray, R. J. (1987). Methodology for case-control studies with prevalent cases. *Biometrika* 74, 191–195.
- Bhatti, P., Doody, M. M., Alexander, B. H., Yuenger, J., Simon, S. L., Weinstock, R. M., et al. (2008). Breast cancer risk polymorphisms and interaction with ionizing radiation among US radiologic technologists. *Cancer Epidemiol Biomark Prev* 17, 2007–2011.
- Chan, K. C. G. (2013). Survival analysis without survival data: Connecting length-biased and case-control data. *Biometrika* 100, 764–770.
- Cheng, Y. J. and Huang, C. Y. (2014). Combined estimating equation approaches for semiparametric transformation models with length-biased survival data. *Biometrics* 70, 608–618.
- Cook, R. J. and Bergeron, P. J. (2011). Information in the sample covariate distribution in prevalent cohorts. *Stat Med* 30, 1397–1409.

- Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* 66, 403–411.
- Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* 85, 619–630.
- Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann Stat* 22, 300–325.
- Ross, S. M. (1996). *Stochastic Processes* (2nd edition). Wiley.
- Schlesselman, J. J. (1982). *Case-control Studies: Design, Conduct, Analysis*. Oxford University Press.
- Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *J Am Stat Assoc* 86, 130–143.
- Zhu, H., Ning, J., Shen, Y., and Qin, J. (2017). Semiparametric density ratio modeling of survival data from a prevalent cohort. *Biostatistics* 18, 62–75.

## SUPPORTING INFORMATION

Web Appendices referenced in Sections 3, 4, and 5 include the definition of  $V$  and calculation of  $\Sigma$ , the proof and regularity conditions for Theorem 1, further simulation results, additional data tables and R code are available at the *Biometrics* website, Wiley Online Library.

**How to cite this article:** Maziarz M, Liu Y, Qin J, Pfeiffer RM. Inference for case-control studies with incident and prevalent cases. *Biometrics*. 2019;75:842–852. <https://doi.org/10.1111/biom.13023>

## Appendix 1: Derivation of the profile log-likelihood (7)

Letting  $w_1(x) = \exp(\alpha^* + x\beta)$  and  $w_2(x) = \exp\{v^* + x\beta + \log \mu(x, \gamma)\}$  where  $\alpha^*$  and  $v^*$  are defined as in equations (2) and (4), respectively, we rewrite the likelihood in (6) as

$$\mathcal{L} = \prod_{i=1}^N f_0(x_i) \prod_{i=n_0+1}^{n_0+n_1} w_1(x_i) \prod_{i=n_0+n_1+1}^N \{w_2(x_i)S(a_i | x_i, \gamma)/\mu(x_i, \gamma)\}.$$

Following Qin (1998), we estimate  $p_i = f_0(x_i) = P(X = x_i)$ ,  $i = 1, \dots, N$ , empirically, accommodating the constraints:  $\sum_{i=1}^N p_i = 1$ ,  $p_i \geq 0$ ;  $\sum_{i=1}^N p_i w_1(x_i) = 1$ , and  $\sum_{i=1}^N p_i w_2(x_i) = 1$ , in the log-likelihood using Lagrange multipliers,  $\lambda_i$ ,  $i = 0, 1, 2$ , resulting in

$$\begin{aligned} \ell_c = & \sum_{i=1}^N \log p_i + \sum_{i=n_0+1}^{n_0+n_1} \log w_1(x_i) \\ & + \sum_{i=n_0+n_1+1}^N \log w_2(x_i) \{S(a_i | x_i, \gamma)/\mu(x_i, \gamma)\} \\ & + \lambda_0 \left(1 - \sum_{i=1}^N p_i\right) \\ & + N \lambda_1 \left\{1 - \sum_{i=1}^N p_i w_1(x_i)\right\} + N \lambda_2 \left\{1 - \sum_{i=1}^N p_i w_2(x_i)\right\}. \end{aligned}$$

Taking derivatives with respect to the  $p_i$ 's we explicitly compute  $\lambda_0$  and  $p_i$ . As

$$\frac{\partial \ell_c}{\partial p_i} = \frac{1}{p_i} - \lambda_0 - N \lambda_1 w_1(x_i) - N \lambda_2 w_2(x_i) = 0 \quad (A.1)$$

and thus  $\sum_{i=1}^N p_i \frac{\partial \ell_c}{\partial p_i} = N - \lambda_0 - N \lambda_1 - N \lambda_2 = 0$ , and  $\hat{\lambda}_0 = N(1 - \lambda_1 - \lambda_2)$ . Plugging  $\hat{\lambda}_0$  into equation (A.1) yields

$$p_i = (N [1 + \lambda_1 \{w_1(x_i) - 1\} + \lambda_2 \{w_2(x_i) - 1\}])^{-1} \quad (A.2)$$

and the profile log-likelihood for the remaining parameters  $\ell_p(\lambda_1, \lambda_2, \alpha^*, v^*, \beta, \gamma)$ , is

$$\begin{aligned} \ell_p(\lambda_1, \lambda_2, \alpha^*, v^*, \beta, \gamma) = & - \sum_{i=1}^N \log [1 + \lambda_1 \{w_1(x_i) - 1\} + \lambda_2 \{w_2(x_i) - 1\}] \\ & + \sum_{i=n_0+1}^{n_0+n_1} \log w_1(x_i) + \sum_{i=n_0+n_1+1}^N \log w_2(x_i) \\ & \times \{S(a_i | x_i, \gamma) / \mu(x_i, \gamma)\} - N \log(N). \end{aligned} \quad (A.3)$$

Differentiation of  $\ell_p(\lambda_1, \lambda_2, \alpha^*, v^*, \beta, \gamma)$  yields

$$\frac{\partial \ell_p}{\partial \lambda_k} = \sum_{i=1}^N \frac{w_k(x_i) - 1}{1 + \lambda_1 \{w_1(x_i) - 1\} + \lambda_2 \{w_2(x_i) - 1\}} = 0, k = 1, 2. \quad (A.4)$$

and

$$\frac{\partial \ell_p}{\partial \alpha^*} = - \sum_{i=1}^N \frac{\lambda_1 w_1(x_i)}{1 + \lambda_1 \{w_1(x_i) - 1\} + \lambda_2 \{w_2(x_i) - 1\}} + n_1 = 0 \quad (A.5)$$

$$\frac{\partial \ell_p}{\partial v^*} = - \sum_{i=1}^N \frac{\lambda_2 w_2(x_i)}{1 + \lambda_1 \{w_1(x_i) - 1\} + \lambda_2 \{w_2(x_i) - 1\}} + n_2 = 0 \quad (A.6)$$

$$\begin{aligned} \frac{\partial \ell_p}{\partial \beta} = & - \sum_{i=1}^N \frac{\lambda_1 x_i w_1(x_i) + \lambda_2 x_i w_2(x_i)}{1 + \lambda_1 \{w_1(x_i) - 1\} + \lambda_2 \{w_2(x_i) - 1\}} \\ & + \sum_{i=n_0+1}^{n_0+n_1} x_i + \sum_{i=n_0+n_1+1}^N x_i = 0 \\ \frac{\partial \ell_p}{\partial \gamma} = & - \sum_{i=1}^N \frac{\lambda_2 w_1(x_i) \frac{\partial}{\partial \gamma} \mu(x_i, \gamma)}{1 + \lambda_1 \{w_1(x_i) - 1\} + \lambda_2 \{w_2(x_i) - 1\}} \\ & + \sum_{i=n_0+n_1+1}^N \frac{\partial}{\partial \gamma} \log S(a_i | x_i, \gamma) = 0. \end{aligned}$$

Next, we solve for  $\lambda_1$  and  $\lambda_2$ . Using equation (A.2) and that  $\sum_{i=1}^N p_i = 1$ , it follows that

$$\sum_{i=1}^N 1 / [1 + \lambda_1 \{w_1(x_i) - 1\} + \lambda_2 \{w_2(x_i) - 1\}] = N \quad (A.7)$$

From equations (A.4) and (A.7) we have

$$\begin{aligned} \sum_{i=1}^N w_k(x_i) / [1 + \lambda_1 \{w_1(x_i) - 1\} + \lambda_2 \{w_2(x_i) - 1\}] = N, \\ k = 1, 2. \end{aligned} \quad (A.8)$$

Then, using (A.5) and (A.8)

$$\begin{aligned} \frac{\partial \ell_p}{\partial \alpha^*} = 0 \Rightarrow \sum_{i=1}^N \frac{\lambda_1 w_1(x_i)}{1 + \lambda_1 \{w_1(x_i) - 1\} + \lambda_2 \{w_2(x_i) - 1\}} \\ = n_1 \Rightarrow \hat{\lambda}_1 = \frac{n_1}{N} \end{aligned}$$

and similarly, from (A.6) and (A.8) for  $v^*$  we get that  $\hat{\lambda}_2 = n_2 / N$ . Plugging the estimates  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  into equation (A.3) yields the profile likelihood in (7).