University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2005 *Paper* 149

Survival Point Estimate Prediction in Matched and Non-Matched Case-Control Subsample Designed Studies

Annette M. Molinaro* Mark J. van der Laan†
Dan H. Moore‡ Karla Kerlikowske**

*Division of Biostatistics, Epidemiology and Public Health, Yale University , annette.molinaro@yale.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

‡Dept. of Epidemiology & Biostatistics, University of California, San Francisco, dmoore@cc.ucsf.edu

**Dept. of Medicine & Dept. of Epidemiology and Biostatistics, University of California, San Francisco, kerliko@itsa.ucsf.edu

http://biostats.bepress.com/ucbbiostat/paper149

Survival Point Estimate Prediction in Matched and Non-Matched Case-Control Subsample Designed Studies

Annette M. Molinaro, Mark J. van der Laan, Dan H. Moore, and Karla Kerlikowske

Abstract

Providing information about the risk of disease and clinical factors that may increase or decrease a patient's risk of disease is standard medical practice. Although case-control studies can provide evidence of strong associations between diseases and risk factors, clinicians need to be able to communicate to patients the age-specific risks of disease over a defined time interval for a set of risk factors.

An estimate of absolute risk cannot be determined from case-control studies because cases are generally chosen from a population whose size is not known (necessary for calculation of absolute risk) and where duration of follow-up is not known (necessary for calculation of incidence). This problem can sometimes be overcome by using a nested case-control design.

We have collected data on a National Cancer Institute funded population-based cohort study. This study contains a matched set of cases and controls within the cohort. This design is more cost-efficient than a full cohort study since expensive predictor variables (genomic measures, sex hormone levels, mammographic breast density) are measured on all of the cases, but on only a sample of the cohort who did not develop the outcome of interest (the controls). In addition, this design avoids the potential biases of conventional case-control studies that draw cases and controls from different populations. Importantly, the presence or absence of the outcome of interest has been established for the entire cohort within the same time period.

The specifics of the sampling in our study do not adhere to the assumptions for absolute risk estimation methods previously developed in the literature. Here we introduce a novel method which provides locally efficient estimators to predict the absolute risk of a cohort from measures only taken on the matched case-control participants. The proposed method is evaluated using simulation studies and survival data from women with ductal carcinoma in situ, a non-invasive form of breast cancer. A generalization of the proposed method is related to other similar sampling designs such as nested case-control, case-cohort, and two-stage case-control.

Contents

1 Introduction

Providing information about the risk of disease and clinical factors that may increase or decrease a patient's risk of disease is standard medical practice. For example, a clinician may inform a 50-year-old woman that her lifetime risk of breast cancer is 1 in 8, but that her risk of breast cancer in the next ten years is 2.5%. If the same 50-year-old woman has mammographically dense breasts, case-control studies suggest she has a relative increase in the risk of breast cancer of 3-fold compared to a 50-year old who does not have dense breasts. Although case-control studies provide evidence of strong associations between diseases and risk factors in the form of relative risks or of odds ratios, clinicians need to assess and express to patients the age-specific absolute risk of disease for a given time period and set of risk factors. Thus, for the 50-year-old woman with mammographically dense breasts, her clinician needs to be able to communicate the absolute risk of breast cancer in the next ten years given her present age and mammographically dense breasts.

Absolute risk is defined as the observed or calculated probability of an event (e.g., occurrence of breast cancer) in the population under study. It can be expressed as a simple probability in a cross-sectional study, or as a hazard or survival function in a longitudinal study. In clinical longitudinal studies, the most frequent events or outcomes of interest are times to initial occurrence of disease, recurrence of disease, and death from disease. In either a cross-sectional or a longitudinal study, if researchers were able to measure all risk factors and observe all patients until the outcome of interest, e.g., recurrence of breast cancer, the absolute risk would be trivial to calculate. However, this is rarely the case. Instead, researchers are faced with two problems. First, at the end of a study, some patients may have dropped out, been lost to follow-up, or not had the particular event of interest. In this situation, the last date of follow-up or date at end of study is recorded and referred to as the censored time to event.

A second problem is that interesting biologic and epidemiologic markers can be time consuming and expensive to collect, store, assess, and analyze. To address this problem different sampling approaches have been implemented. In the nested case-control study design an eligible cohort has been recruited, and subsequently the participants with the event of interest, e.g., recurrence of breast cancer, are identified and labeled as cases (Mantel, 1973; Lidel et al., 1977; Breslow, 1996). At the time of each individual case's event of interest a sample of those without current or previous events are

selected and labeled as controls. This allows controls to be selected for more than one case and cases still at risk to serve as controls for those cases who have had preceding failures. A second sampling scheme is the case-cohort design (Prentice, 1986). In this design a subcohort is randomly selected from the entire cohort in addition to all cases occurring outside the subcohort allowing a comparison group for all failure times. The third sampling is a more general setting for the nested case-control: the two-stage case-control design. The first stage is the recruitment and collection of data including the outcome of a cohort; the second stage entails selecting subsamples of the cases and controls from the first stage to assess the additional desired covariates (Breslow and Cain, 1988). All three designs allow researchers the advantage of only collecting the interesting biologic and epidemiologic markers on those participants in the cohort assigned to the subsample (i.e., the nested, subcohort, or stage two group) (Ernster, 1994; Wacholder, 1991). Once the markers, or risk factors, are determined in the subsample, measures such as the relative risk (RR) and odds ratio (OR) can be evaluated and extrapolated to the entire cohort. Methods for estimating the RR and OR have been thoroughly explored for each of the three sampling designs (Goldstein et al., 1992; Wacholder and Weinberg, 1994; Self and Prentice, 1988; Breslow and Zhao, 1988; Flanders and Greenland, 1991; Zhao and Lipsitz, 1992).

Other measures, such as the absolute risk of recurrence, are not as easy to extrapolate to the entire cohort. Naive estimates of the absolute risk can be skewed because the proportion with disease (i.e., cases) is higher in the subsample than in the cohort. In nested case-control sampling designs, numerous advances have been made in the last decade based on the known sampling probabilities from the cohort (Goldstein et al., 1992; Benichou and Gail, 1990, 1995; Langholz and Goldstein, 1996; Borgan et al., 1995; Borgan and Langholz, 1993; Lanholz and Borgan, 1997). These methods use relative risk estimates from the nested case-control study as well as assume a semi-parametric model, i.e., the Cox Proportional Hazards model, to assess the risk for the cohort. The same model assumption is made when assessing the absolute risk in case-cohort studies (Self and Prentice, 1988). In two-stage case-control studies incidence estimation has been explored with a Poisson pseudo-likelihood approach (Benichou et al., 1997). The suggested absolute risk procedure for the two-stage design assumes the same Cox Proportional Hazards model and risk set sampling as in the nested case-control design (Langholz and Goldstein, 1996).

We have collected data on a National Cancer Institute funded population-

based cohort study. In this study, data were collected on 1036 women aged 40 and older diagnosed with ductal carcinoma *in situ* (DCIS) from $1983-94$ treated by lumpectomy alone. The event of interest is disease recurrence, defined as DCIS or invasive breast cancer diagnosed in the ipsilateral breast of the initial DCIS lesion or at a distant site more than 6 months following initial diagnosis and treatment of DCIS. The purpose of this study was to identify epidemiological and histological variables associated with recurrence. A detailed description can be found in Kerlikowske et al. (Kerlikowske et al., 2003).

Epidemiological variables were collected on all participants in the cohort. However, due to constraints on tissue collection and the cost of certain biological markers, the histological variables were collected only in the matched case-control study (Figure 1). The matching variable, year of diagnosis, was chosen with the intent of insuring an equivalent time for cases *and* controls to recur. Due to this matching, the assumption that only at a case's failure time are the controls chosen is not valid. Thus, these data deviate from the assumptions made in the three sampling designs and corresponding absolute risk estimation methods. The goal of this paper is to explore a *new* method which addresses the aforementioned question of estimating the absolute risk as a missing data problem. The proposed approach is substantially different from assuming a parametric or semi-parametric model, inheriting the estimated relative risk, and/or making the previously mentioned assumptions on the risk set. Additional interests in a new method are: to account for informative censoring, which has previously not been addressed in this setting; to develop locally efficient point estimators of the survival distributions; and, to construct an estimator which accommodates partial data (e.g., censored time to event and missing measurements on variables only collected in the subsample) that reduces to the proper estimator when all data are available.

2 General Methodology

Case-cohort, nested case-control, and two-stage case-control studies are set up to capture extensive information provided by all (or most) of the subjects with events in a cohort, subsequently labeled as cases, as well as a sample of the subjects with non-events, subsequently labeled as controls. In a generalization of these study designs, there are two types of missingness: the

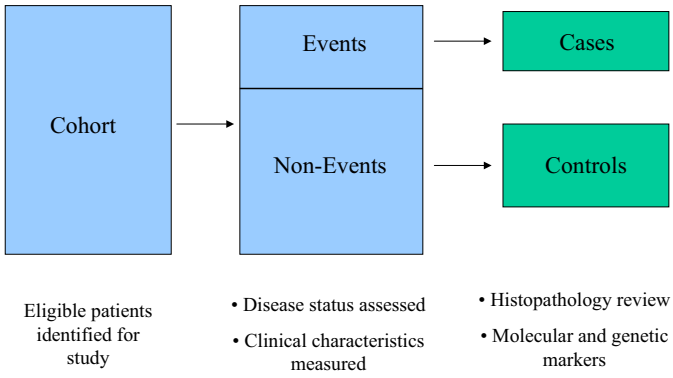

Figure 1: *Depiction of the NCI Cohort with Case-Control Subsample Study Design.* Once disease status is ascertained on eligible patients assignment into the case-control study is made and valuable markers are assessed.

censored observations (i.e., the non-event's follow-up time) and the missing measurements for the variables collected only within the subsample and not on the entire cohort. We refer to the entire cohort with *all* information collected as the *full data world* (Section 2.1) and that which has the aforementioned two levels of missingness as the *observed data world* (Section 2.2). A question of interest in this setting is how to estimate absolute risk of the cohort based on those variables measured only within the subsample.

It is the purpose of this manuscript to derive estimators of the cohort's absolute risk that link the full and observed data worlds with the following two requirements. First, when applied to complete data, the observed data methodology should reduce to the full data methodology. Second, we wish to incorporate external (to the estimator) covariate processes to allow for informative censoring and a gain in efficiency. In Section 2.3, we propose to use the general estimating function methodology of van der Laan and Robins (2002) to map the full data estimating function into an observed data estimating function having the same expected value and leading to an efficient estimator.

2.1 Full data structure

In the *full data world*, suppose one observes n independent and identically distributed (i.i.d) observations, $X_1, \ldots, X_n$, of a full data structure $X = (T, E, E^*)$. Let T denote the event time. This event could be initial occurrence of a disease, recurrence of a disease, or death from disease. E and E^* denote baseline covariates. These two types of baseline covariates designate those measured when a patient enters the study, E, and those measured upon assignment to the subsample, E^*. Although E^* are collected subsequent to E they are measurements on materials collected at baseline, e.g., histopathologic evaluation of nuclear grade for the initial tumor. In the DCIS study all E^* are discrete, e.g., nuclear grade of high, medium, or low. As such, for the purposes of this discussion E^* is discrete. Denote the distribution of the full data structure X by F_X. Our parameter of interest, $\vartheta(t, \delta)$, is the probability of no event up to time t given that the covariate of interest E^* is equal to a specific value δ, i.e.,

$$\vartheta(t, \delta) = Pr(T > t \mid E^* = \delta) = \frac{E(I(T > t, E^* = \delta))}{E(I(E^* = \delta))} = \frac{\mu(t, \delta)}{\mu(\delta)}, \qquad (1)$$

where $\delta \in \{0, 1, \cdots\}$, and we can estimate $\mu(t, \delta)$ and $\mu(\delta)$ with the full data estimating functions $I(T > t, E^* = \delta)$ and $I(E^* = \delta)$, respectively. The absolute risk for time t and $E^* = \delta$ is equal to $1 - \vartheta(t, \delta)$.

2.2 Observed data structure

In the *observed data world*, we rarely have measurements for all of the relevant variables (e.g., E^*, T) in the full data structure. Here we are confronted with both missing covariate values for E^* and missing event times. For the former, E^* is measured on the subsample participants, whereas for the cohort members excluded from the subsample no E^* is available. For the latter, we observe, $\tilde{T}$, the minimum of the event time T and a univariate censoring variable C, i.e., $\tilde{T} = min(T, C)$. This missing, or *censored*, event data can be due to drop out, C^F, or the end of follow-up in a study, C^*. Here, we let C denote the minimum of the two, i.e., $C = min(C^F, C^*)$. By convention, if T occurs prior to C, we set $C = \infty$; thus, C is *always* observed.

In the observed data world, suppose one observes n i.i.d. observations, $O_1, \ldots, O_n$, of the *observed data structure*,

$$O = (C, \tilde{T}, \Delta, E, \Phi E^*, \Phi),$$

where $\Phi = I(E^* \text{ is measured})$ and $\Delta = I(C > T)$. The random variable O has a distribution indexed by the full data distribution, F_X, and the conditional bivariate distribution, $G(\cdot \mid X)$, of the censoring variable C and missingness variable Φ given X, i.e., $O_i \sim P = P_{F_X,G}$. Let the empirical distribution of $O_1, \ldots, O_n$ be denoted by P_n. $G(\cdot \mid X)$ is referred to as the *censoring* or *coarsening mechanism*. The *survivor function* for the censoring mechanism is denoted by $\bar{G}(c \mid X) = Pr(C \geq c \mid X)$.

An assumption we will make on the censoring mechanism is coarsening at random (CAR). Detailed descriptions of CAR can be found in Gill et al. (1997), van der Laan and Robins (2002) (Section 1.2.3, in particular), and Robins and Rotnitzky (1992). For this particular observed data structure and definition of full data structure, CAR is equivalent to:

$$g(c, \phi \mid X) = Pr(C = c, \Phi = \phi \mid T, E, E^*) = m(c, \phi, I(c > T), E, \phi E^*)$$

for some measurable function m. We have the following factorization for $g(c, \phi \mid X)$:

$$Pr(C = c, \Phi = \phi \mid T, E, E^*) = Pr(\Phi = \phi \mid C = c, T, E, E^*) \times Pr(C = c \mid T, E, E^*). \quad (2)$$

Thus CAR holds if the censoring density $g(c, \phi \mid X)$ is only a function of the observed data O. We shall assume that the Lebesgue hazard $\lambda_C(t \mid T, E, E^*)$ corresponding to the censoring mechanism given the full data is only a function of E, i.e., $\lambda_C(t \mid T, E, E^*) = \lambda_C(t \mid E)$ for $t < T$.

Additionally, we shall assume that $Pr(\Phi = \phi \mid C = c, T = t, E, E^*) = m(I(c > t), E)$ and, thus, the missingness of E^* is solely a function of the always observed censoring time (as noted above if $C > T$ then $C = \infty$) and the baseline covariates E collected on all members of the cohort. There are two scenarios which pertain to this assumption: when all recurrences are included as cases and when only a random sample of the recurrences are counted as cases. In the following, we provide examples and explanations of both scenarios.

Scenario One: All recurrences are cases In this scenario, all participants in the cohort who have an event (e.g., recurrence of disease) have probability 1 of having E^* measured. Those without an event, i.e., the censored participants, have a probability of E^* being measured equal to a function of E. This can be written as:

$$\Pi(\Phi = 1 \mid C = c, T, E, E^*) = I(c > T) + I(c < T) \times \rho_2(E), \quad (3)$$

where $\rho_2(E)$ is a function of the always observed baseline covariates E.

Allotting a function of E allows us to include unmatched and matched case-control study designs, varying levels of matching, e.g., one control to one case or two controls to one case, and the case-cohort study design. For example, in an *unmatched* one-to-one case-control study $\rho_2(E)$ might be a constant, i.e., the probability of non-recurrences chosen for controls equals the number of cases divided by the number of non-recurrences. In a case-cohort study design, $\rho_2(E)$ might also be a constant defined by the desired number in the subcohort divided by the number of non-recurrences in the full cohort.

In a *matched* case-control study, patients are typically matched on a variable with an effect for which the study designers wish to control. We denote this variable as E_{match} and note that it is an element of E. The matching can be one case to one control $(1:1)$, one case to two controls $(1:2)$, one case to three controls $(1:3)$, etc. In $1:j$ matching, a recurrence with $E_{match} = e$ is matched to j non-recurrence(s) from the pool of all non-recurrences with the value e for E_{match}. Thus, the probability of a non-recurrence having E^* measured (i.e., being chosen as a control) is a function of E_{match} as well as the number of recurrences with the same value, thus a function of E_{match}.

Scenario Two: Random sample of recurrences are cases A second scenario is when not all of the participants with events have E^* measured. For example, this may occur when the pathologist has not finished reading all of the slides or when for purposes of cost-effectiveness not all recurrences will be measured. In this situation the probability of E^* being measured given that the person had an event is not equal to 1. One way to write this is:

$$\Pi(\Phi = 1 \mid C = c, T, E, E^*) = I(c > T) \times \rho_1 + I(c < T) \times \rho_2(E), \quad (4)$$

where $\rho_2(E)$ is a function of the always observed baseline covariates and ρ_1 is a constant, e.g., the proportion of recurrences with E^* measured of the total number of recurrences.

For either *Scenario One* or *Two*, we can define $\rho_2(E) = \frac{kP(\Delta=1|E)}{P(\Delta=0|E)}$, where k is the number of controls for every case, E represents the always observed baseline covariates (Note: as above, any matching variable E_{match} is in the set E), and $\Delta = 1$ for cases and $\Delta = 0$ for controls.

In the following section we outline how to map the full data world as described in Section 2.1 into the observed data world as described in Section 2.2 for the purpose of estimating the parameter of interest, $\vartheta(t, \delta)$.

2.3 Mapping the full data world to the observed data world

Our stated goal is to find an estimator of the parameter of interest, $\vartheta(t, \delta)$, the probability of no event up to time t given that $E^* = \delta$. In the full data world, where we have both T and E^* on every participant in the cohort, we can use full data estimating functions for this parameter. In the numerator of Equation 1, we write $\mu(t, \delta)$ as the expected value of a full data estimating function, i.e., $\mu(t, \delta) = E_{F_X}[D_{(t,\delta)}(X)]$, where,

$$D_{(t,\delta)}(X) = I(T > t, E^* = \delta). \tag{5}$$

Similarly, for the denominator of Equation 1, we write $\mu(\delta)$ as the expected value of a full data estimating function, i.e., $\mu(\delta) = E_{F_X}[D_\delta(X)]$, where,

$$D_{(\delta)}(X) = I(E^* = \delta). \tag{6}$$

However, in the observed data world, we are faced with censoring on the event time and missingness on E^*, both affecting the numerator, $\mu(t, \delta)$, and denominator, $\mu(\delta)$, of our parameter of interest. In order to address the censoring and missingness we must replace the full data estimating functions above with observed data estimating functions which have the same expected value.

The general estimating function methodology of van der Laan and Robins (2002) can be used expressly for this purpose. Specifically, the methodology allows full data estimating functions, $D(X)$, to be mapped into observed data estimating functions, $IC(O \mid Q, G, D)$ (Note: IC is an abbreviation for *influence curve* as denoted in van der Laan and Robins (2002)), indexed by *nuisance parameter* G and, possibly, $Q = Q(F_X)$. Furthermore, the observed data estimating functions have the same expected value as the full data estimating function, i.e.,

$$E_P[IC(O \mid Q_0, G_0, D)] = E_{F_X}[D(X)] \qquad \text{if } G_0 = G \text{ or } Q_0 = Q(F_X).$$

There are several candidates of the mapping $D(X) \rightarrow IC(O \mid Q, G, D)$; in particular, we are interested in the inverse probability of censoring weighted

and the doubly robust inverse probability of censoring weighted estimating functions. In the following section we outline the first and refer the reader to descriptions of the latter. For the interested reader thorough descriptions of both can be found in van der Laan and Robins (van der Laan and Robins, 2002).

2.4 Inverse probability of censoring weighted estimating function.

The *inverse probability of censoring weighted* (IPCW) estimating function was introduced by Robins and Rotnitzky (1992). Its name derives from the fact that the full data estimating function $D(X)$ is weighted by the inverse of a censoring probability. The general IPCW estimating function is written as:

$$D(X)\frac{I(X \text{ is observed})}{P(X \text{ is observed} \mid X)},$$

Thus, the IPCW estimating equation which accounts for censoring and the missingness on E^* for $\mu(t,\delta)$ in Equation 1 is:

$$IC_0^{\mu(t,\delta)}(O \mid G, D) = D_{(t,\delta)}(X)\frac{I(\Delta = 1, \Phi = 1)}{Pr(\Delta = 1, \Phi = 1 \mid X)}, \tag{7}$$

where $D_{(t,\delta)}(X)$ is defined in equation 5, $\Phi = I(E^* \text{ is measured})$, and Δ is the event indicator, i.e., $\Delta = I(C > T)$.

Scenario One As described in *Scenario One* of Section 2.2, every participant who has an event (subsequently defined as a case) also has E^* measured $\Phi = 1$. Thus, we can replace the denominator of Equation 7 with $Pr(\Delta = 1 \mid X)$. Now given

$$E[\Delta\Phi \mid X] = Pr(\Delta = 1, \Phi = 1 \mid X) = \bar{G}(T \mid E) > 0, \qquad F_X\text{-a.e.},$$

the IPCW estimating function can be shown to have the same expected value as the full data estimating function:

$$\begin{aligned} E\left[\frac{D_{(t,\delta)}(X)I(\Delta = 1, \Phi = 1)}{\bar{G}(T \mid E)}\right] &= E\left[E\left[\frac{D_{(t,\delta)}(X)I(\Delta = 1, \Phi = 1)}{\bar{G}(T \mid E)} \mid X\right]\right] \\ &= E\left[\frac{D_{(t,\delta)}(X)E[\Delta\Phi \mid X]}{\bar{G}(T \mid E)}\right] \\ &= E\left[D_{(t,\delta)}(X)\right] = \mu(t,\delta). \end{aligned}$$

Scenario Two In *Scenario Two* of Section 2.2, not every patient who has an event has E^* measured, instead there is a random sampling of subjects with events that become cases. Given the example in Equation 4 for $Pr(\Phi = \phi \mid C = c, T, E, E^*)$, we can write:

$$\begin{aligned}
&Pr(\Phi = 1, C \geq T \mid T, E, E^*) \\
&= \int_{c=T}^{\infty} Pr(\Phi = 1, C = c \mid X) \\
&= \int_{c=T}^{\infty} Pr(\Phi = 1 \mid C = c, X) Pr(C = c \mid X) \\
&= \int_{c=T}^{\infty} [I(T < c)\rho_1 + I(T > c)\rho_2(E, \Delta)] Pr(C = c \mid X) \\
&= \int_{c=T}^{\infty} I(T < c)\rho_1 Pr(C = c \mid X) + I(T > c)\rho_2(E, \Delta) Pr(C = c \mid X) \\
&= \int_{c=T}^{\infty} \rho_1 \times Pr(C = c \mid X) \\
&= \rho_1 \times \bar{G}(T \mid E)
\end{aligned}$$

Thus, the denominator of Equation 7 can be replaced with $\rho_1 \times Pr(\Delta = 1 \mid X)$, where ρ_1 is a constant. In this scenario,

$$E[\Delta\Phi \mid X] = Pr(\Delta = 1, \Phi = 1 \mid X) = \rho_1 \times \bar{G}(T \mid E) > 0, \qquad F_X\text{-a.e.},$$

and thus we can show that

$$E\left[\frac{D_{(t,\delta)}(X) I(\Delta = 1, \Phi = 1)}{\rho_1 \times \bar{G}(\cdot \mid E)}\right] = E\left[D_{(t,\delta)}(X)\right] = \mu(t, \delta),$$

in the same fashion as in *Scenario One*.

The fundamental procedure as illustrated in *Scenario One* and *Two* suggests the utility of the IPCW observed data estimating function, $IC_0^{\mu(t,\delta)}(O \mid G, D_{(t,\delta)}(X) = I(T > t, E^*))$, with nuisance parameter G. For Scenario One, the corresponding numerator of the parameter of interest in Equation 1 is the empirical mean:

$$\hat{\mu}(t, \delta) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(T_i > t, E_i^* = \delta) I(\Delta_i = 1, \Phi_i = 1)}{\bar{G}_n(T \mid E_i)}, \tag{8}$$

where $\bar{G}_n$ is an estimator of the nuisance parameter $\bar{G}$. One can estimate the nuisance parameter $\bar{G}$ in the IPCW using any of the covariates in E in

order to allow for informative censoring and a gain in efficiency. (Note: For *Scenario Two* the denominator in Equation 8 is replaced by $\rho_1 \times \bar{G}_n(T \mid E)$).

Equation 8 is shown with a 'global' indicator of event, i.e., $\Delta = \Delta^{global} = I(C > T)$; however, one may be interested in a time specific indicator of event, e.g., $\Delta^t = I(C > t)$, where t is the time of interest. Then, in Equation 8, $\bar{G}_n(T \mid E_i)$ is replaced by $\bar{G}_n(t \mid E_i)$, and $I(\Delta_i = 1, \Phi_i = 1)$ is replaced by $I(\Delta_i^t = 1, \Phi_i = 1)$. Unless otherwise indicated, Δ refers to Δ^{global} in this manuscript.

If a Cox proportional hazards model is assumed for the censoring mechanism G, then

$$\lambda_C(t \mid E) = \lambda_0(t) \exp(\beta^T E),$$

where E is the set of always observed baseline covariates. Standard software can then be employed to obtain maximum (partial) likelihood estimators of the baseline hazard function λ_0 and the regression coefficients β (e.g., `coxph` function in R). Importantly, the estimator $\bar{G}_n(T \mid X)$ and $I(\Delta = 1, \Phi = 1)$ are functions of $O = (C, \Delta, E, \Phi E^*, \Phi)$ and thus the resulting estimator $\hat{\mu}(t, \delta)$ depends only on the observed data structure, $O_1, \ldots, O_n$.

An IPCW estimating function which solely accounts for the missingness on E^* can be similarly built for the full data estimating function in Equation 6. This observed data estimating function which corresponds with the denominator, $\mu(\delta)$, in equation 1 is:

$$IC_0^{\mu(\delta)}(O \mid G, D_{(\delta)}) = D_{(\delta)}(X) \frac{I(\Phi = 1)}{Pr(\Phi = 1 \mid X)}, \tag{9}$$

where $\Phi = I(E^*$ is measured). Thus, the estimator for the denominator of the parameter of interest is the empirical mean:

$$\hat{\mu}(\delta) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(E_i^* = \delta) I(\Phi_i = 1)}{P(\Phi_i = 1 \mid E_i)}, \tag{10}$$

A simple logistic model can be used to estimate $P(\Phi_i = 1 \mid E_i)$. Similar to $\hat{\mu}(t, \delta)$, all components of $\hat{\mu}(\delta)$ are functions of O and thus $\hat{\mu}(\delta)$ also only depends on the observed data structure. Given Equations 8 and 10, it follows that $\vartheta(t, \delta)$ can be estimated by:

$$\hat{\vartheta}(t, \delta) = \frac{\hat{\mu}(t, \delta)}{\hat{\mu}(\delta)} = \frac{\frac{1}{n} \sum_{i=1}^{n} \frac{I(T_i > t, E_i^* = \delta) I(\Delta_i = 1, \Phi_i = 1)}{\bar{G}_n(\cdot \mid E_i)}}{\frac{1}{n} \sum_{i=1}^{n} \frac{I(E_i^* = \delta) I(\Phi_i = 1)}{P(\Phi_i = 1 \mid E_i)}}, \tag{11}$$

which only depends on the observed data structure O. In *Scenario Two*, the denominator of the numerator in Equation 11 is replaced by $\rho_1 \times \bar{G}_n(\cdot \mid E)$. As also noted above, if $\Delta = \Delta^{global}$, then $\bar{G}_n(\cdot \mid E_i) = \bar{G}_n(T_i \mid E_i)$; whereas, if $\Delta = \Delta^t$, then $\bar{G}_n(\cdot \mid E_i) = \bar{G}_n(t \mid E_i)$. Conditions for the IPCW estimating functions to be consistent are that $\bar{G}(\cdot \mid X) > \epsilon > 0$, F_X-a.e., for some $\epsilon > 0$, and that $\bar{G}_n$ is a consistent estimator for $\bar{G}$.

In practice the numerator and denominator are estimated separately; therefore, the ratio may need to be weighted guaranteeing that at time 0 the probability of no event is equal to 1, i.e., $\hat{\vartheta}(0, \delta) = 1$. This can be achieved by dividing the ratio in Equation 11 by the parameter of interest evaluated at 0, i.e., $\frac{\hat{\vartheta}(t,\delta)}{\hat{\vartheta}(0,\delta)}$. We shall denote this weighted estimator of the parameter of interest as $\hat{\vartheta}_w(t, \delta)$. Interestingly, this estimator results in the following:

$$\hat{\vartheta}_w(t, \delta) = \frac{\hat{\vartheta}(t, \delta)}{\hat{\vartheta}(0, \delta)} = \frac{\frac{\hat{\mu}(t,\delta)}{\hat{\mu}(\delta)}}{\frac{\hat{\mu}(0,\delta)}{\hat{\mu}(\delta)}},$$

and since $\hat{\mu}(\delta)$ is a constant this new estimator becomes,

$$\hat{\vartheta}_w(t, \delta) = \frac{\hat{\vartheta}(t, \delta)}{\hat{\vartheta}(0, \delta)} = \frac{\hat{\mu}(t, \delta)}{\hat{\mu}(0, \delta)} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{I(T_i > t, E_i^* = \delta) I(\Delta_i = 1, \Phi_i = 1)}{\bar{G}_n(T \mid E_i)}}{\frac{1}{n}\sum_{i=1}^{n} \frac{I(T_i > 0, E_i^* = \delta) I(\Delta_i = 1, \Phi_i = 1)}{\bar{G}_n(T \mid E_i)}}. \tag{12}$$

Now any misspecification of $\mu(\delta) = Pr(E^* = \delta)$ is *not* relevant as it is not included in this weighted estimator of the parameter of interest, $\hat{\vartheta}_w(t, \delta)$. Equation 12 is the new estimator for *Scenario One*; interestingly, in *Scenario Two* one can see that the constant ρ_1 will divide out. Thus, the two estimators become **identical**. This means that Equation 12 can be used as the IPCW estimating function in unmatched as well as matched (including different matching ratios) case-control subsample studies to evaluate the absolute risk at time t given $E^* = \delta$ for the entire cohort. The bootstrap estimate of the standard error for the proposed weighted IPCW estimator can be used to assess the variability and build confidence intervals.

2.5 Doubly robust inverse probability of censoring weighted estimating function.

A second candidate for the mapping $D(X) \rightarrow IC(O \mid Q, G, D)$ is the *doubly robust inverse probability of censoring weighted* (DR-IPCW) estimating function. This function is indexed by two nuisance parameters, G and $Q = Q(F_X)$. An important consequence of CAR is the factorization of the likelihood into an F_X part and a G part. This allows consistency of a maximum likelihood estimator under a model for F_X which only relies on the correct specification of that model.

Given a full data estimating equation D, we let $IC_0(O \mid G_1, D)$ represent an initial estimating function which is unbiased under the observed data distribution $P = P_{F_X,G_1}$. Now we consider the orthogonalized estimating function obtained by subtracting from $IC_0(O \mid G_1, D)$ a projection on a tangent space $T_{CAR}(P_{F_X,G_1})$ of G at P_{F_X,G_1} under a convex model for G. This can be written as:

$$IC(O \mid Q, G_1, G_2, D) \;=\; IC_0(O \mid G_1, D) - \Pi(IC_0(O \mid G_1, D) \mid T_{CAR}),$$

where T_{CAR} can represent the tangent space of any convex model for $g(c, \phi \mid X)$ satisfying the CAR assumption. In particular, it can represent the tangent space for the model $\mathcal{G}(CAR)$ for g only assuming CAR. We propose one such model based on the factorization of g into two mechanisms (censoring and missingness, Equation 2) by separately making assumptions on each of these two mechanisms: namely, $g(c|X)$ is only a function of E for $c < T$, and $g(\phi \mid C, T, E, E^*)$ is only a function of $I(C > T)$ and E. The tangent space of the corresponding submodel of $\mathcal{G}(CAR)$ is convex and it is an orthogonal sum (by factorization of density $g(c, \phi \mid X)$) of two tangent spaces $T_{CAR,1}$ and $T_{CAR,2}$ corresponding with the tangent spaces of missingness and censoring mechanism (two factors in factorization of $g(c, \phi \mid X)$. This can be written as the sum of two projections:

$$\begin{aligned}\Pi\left[IC_0(O \mid G_1, D) \mid T_{CAR}\right] \;=\; & \Pi\left[IC_0(O \mid G_1, D) \mid T_{CAR,1}\right] \\ & + \Pi\left[IC_0(O \mid G_1, D) \mid T_{CAR,2}\right],\end{aligned}$$

where, the projection on T_{CAR1} is:

$$\begin{aligned}&\Pi\left[IC_0(O \mid G, D) \mid T_{CAR_1}\right] = \\ &\quad E[IC_0(O \mid G, D) \mid \Phi, C, E] - E[IC_0(O \mid G, D) \mid C, E]\end{aligned}$$

and the projection on T_{CAR2} as:

$$\Pi\left[IC_0(O \mid G, D) \mid T_{CAR_2}\right] = -\int E[IC_0(O \mid G, D) \mid T > U, C > U, E] dM_{G_2}(U).$$

With $\Delta = \Delta^{global}$, we can write the observed data DR-IPCW estimating equation for equation 5 as:

$$\begin{aligned} IC_0^{\mu(t,\delta)}(O \mid Q, G_1, G_2, D) = \\ \frac{D(X)I(\Delta = 1, \Phi = 1)}{\bar{G}_2(T \mid E)} - \Bigg(E_{Q,G}\left[\frac{D(X)I(\Delta = 1, \Phi = 1)}{\bar{G}_2(T \mid E)} \mid \Phi, C, E\right] \\ - \int_\phi E_{Q,G}\left[\frac{D(X)I(\Delta = 1, \Phi = 1)}{\bar{G}_2(T \mid E)} \mid \Phi = \phi, C, E\right] P_{G_1}(\Phi = \phi \mid C, E)\Bigg) \\ + \int E_{Q,G}\left[\frac{D(X)I(\Delta = 1, \Phi = 1)}{\bar{G}_2(T \mid E)} \mid min(T, C) > U, E\right] dM_{G_2}(U) \end{aligned}$$

Similarly an observed data DR-IPCW estimating function can be written for Equation 6. The reader is directed to van der Laan and Robins (2002) (especially Section 1.6) for a complete discussion and proofs.

It is important to note that if Q is misspecified, this orthogonalized estimating function is unbiased as long as G_1 is correct. This is true because the projection operator under P still maps into functions which have conditional mean zero given X with respect to G_1. The *doubly-robust* name derives from the fact that if Q is correctly specified, however, G is not, the unbiasedness of this estimating function still holds.

3 Simulations

To evaluate our proposed weighted IPCW estimator $\vartheta_w(t, \delta)$ (Equation 12) we present the following results. The intention of this simulation study is to evaluate $\vartheta_w(t, \delta)$'s performance with varying levels of censoring and matching in data which emulates the real data in the DCIS study.

Simulation model for full and observed data structures. The full and observed data structures were chosen to represent a 'real-world' situation as modeled by the DCIS data described in Section 1 and analyzed in Section 4. The full data structure was simulated as $T \mid E^* \equiv Weibull(\text{shape} = \alpha, \text{scale} = \beta_1) \times E^* + Weibull(\text{shape} = \alpha, \text{scale} = \beta_2) * (1 - E^*)$, where $E^* \sim Bernoulli(p_{E^*})$, p_{E^*} is a user-defined probability of having $E^* = 1$, α

is the shape parameter for the Weibull distribution, β_1 is the scale parameter for those observations with $E^* = 1$, and β_2 is the scale parameter for those observations with $E^* = 0$. Additionally, $E \sim Bernoulli(p_E)$ and $E_{match} \sim Multinomial([1 : m], p_M)$, where m, p_E, and p_M are user-specified. For the purposes of the simulations shown in Sections 3.1 and 3.2: $p_{E^*} = 0.38$, $p_E = 0.21$, $m = 10$, and $p_M = 0.1$. These values were chosen to best emulate the real data described in Section 4.

Censoring times C were simulated using a mixture of Weibull and uniform distributions. The mixing proportions for the distributions, as well as the parameters described above (i.e., α, β_1, and β_2), were fine-tuned to achieve a desired level of censoring while ensuring that $Pr(\bar{G}_0(T|W) > 0.1) = 1$, a condition for the IPCW method (Section 2.4). For the simulations, the end of study was set at 180 months, resulting in $T^* = min(T, 180)$, $\tilde{T} = min(T^*, C)$, and the indicator of event defined as $\Delta = I(T^* \leq C)$.

3.1 Unmatched Case-Control Study

In the first simulation study, we generated an unmatched case-control study. In this design every recurrence was included as a case and for each of the recurrences one non-recurrence was randomly assigned as a control. For this design, $P(\Phi = 1 \mid \Delta, E, T, C) = I(C > T) + I(C < T)\rho_2(E)$. Censoring percentages of 70% and 80% were simulated over three sample sizes: 500, 1000, and 5000. For each of the simulations, the proposed weighted IPCW estimator $\vartheta_w(t, \delta)$ was estimated. This estimate was compared to the truth, i.e., the data as simulated in the full data structure, by calculating the mean squared error (MSE). The estimate was assessed at 5, 10, and 15 years for both those at high risk, i.e., $E^* = 1$, and low-risk, i.e., $E^* = 0$ for $B = 100$ bootstrap samples. The average of the bootstrap estimates, average MSE, and the IPCW estimator's standard deviation over the B bootstrap samples are displayed in Table 1. The proposed weighted IPCW estimator is quite accurate as measured by its proximity to the truth. The distance between the two is more apparent by 15-years due to the sparsity in data by the end of the study. The actual survival curves further elucidate the estimator in relation to the truth. In Figure 2, it is apparent how well the IPCW method does in finite samples (n=500).

Table 1: *Simulation Study 3.1.* The cohort's absolute risk of having an event by time t for $E^* = 0$ and $E^* = 1$ based on an unmatched case-control study over 2 censoring levels (column 1), 3 sample sizes (column 2), and 3 time points (columns 4-9). Estimates for the IPCW method (Std. error in parens), the truth, and MSE are based on an average of 100 bootstrap samples.

			5 years		10 years		15 years	
Cens	n	Method	$E* = 0$	$E* = 1$	$E* = 0$	$E* = 1$	$E* = 0$	$E* = 1$
70%	500	IPCW	0.169(.04)	0.322(.07)	0.31(.07)	0.454(.08)	0.332(.07)	0.594(.1)
		Truth	0.153	0.35	0.275	0.528	0.334	0.667
		MSE	.002	.005	.006	.012	.005	.015
	1000	IPCW	0.142(.02)	0.369(.05)	0.245(.03)	0.541(.07)	0.363(.05)	0.64(.08)
		Truth	0.144	0.339	0.247	0.518	0.336	0.641
		MSE	.004	0	.005	.001	.006	.002
	5000	IPCW	0.166(.03)	0.335(.08)	0.246(.04)	0.493(.12)	0.389(.08)	0.604(.13)
		Truth	0.144	0.339	0.247	0.518	0.336	0.641
		MSE	.006	.001	.015	.002	.017	.01
80%	500	IPCW	0.161(.04)	0.312(.07)	0.226(.04)	0.503(.11)	0.32(.07)	0.503(.11)
		Truth	0.153	0.35	0.275	0.528	0.334	0.667
		MSE	.006	.001	.012	.004	.038	.005
	1000	IPCW	0.166(.03)	0.335(.08)	0.246(.04)	0.493(.12)	0.389(.08)	0.604(.13)
		Truth	0.144	0.339	0.247	0.518	0.336	0.641
		MSE	.006	.001	.015	.002	.017	.01
	5000	IPCW	0.128(.01)	0.312(.02)	0.236(.01)	0.488(.03)	0.333(.02)	0.64(.03)
		Truth	0.131	0.293	0.247	0.488	0.338	0.608
		MSE	.001	0	.001	0	.002	0

3.2 Matched Case-Control Study

In the second simulation study, we generated a matched case-control study based. The matching here was based on that of the DCIS study where matching was done on year of diagnosis solely to allow an equivalent follow-up time for potential recurrence. In this simulation each of the recurrences were included as cases. For each of the cases one non-recurrence was assigned as a control based on having the same value of E_{mat}. For this design, $P(\Phi = 1 \mid \Delta, E, T, C) = I(C > T) + I(C < T)\rho_2(E, \Delta)$, where $E_{mat} \in E$. Censoring percentages of 70% and 80% were simulated over three sample sizes: 500, 1000, and 5000. For each of the simulations, the proposed weighted IPCW estimator $\vartheta_w(t, \delta)$ was estimated and compared to the truth, i.e., the data as simulated in the full data structure, by calculating the MSE. The estimate was assessed at 5, 10, and 15 years for both those at high risk, i.e., $E^* = 1$, and low-risk, i.e., $E^* = 0$ for $B = 100$ bootstrap samples. The average of the bootstrap estimates, average MSE, and the IPCW estimator's standard deviation over the B bootstrap samples are displayed in Table 2. The proposed weighted IPCW estimator is quite accurate as measured by its proximity to the truth. Again, the distance between the two is more apparent by 15-years due to the sparsity in data by the end of the study. The actual survival curves further elucidate the estimator in relation to the truth. In Figure 2, it is apparent how well the IPCW method does in finite samples (n=500).

4 Data Analysis

Once the simulation studies validated the proposed estimator's assessment of the absolute risk, we implemented the estimator in a real data set. As described in the Introduction, this data analysis is based on a National Cancer Institute funded population-based cohort study. Extensive details and initial analysis can be found in Kerlikowske et al. (2003).

To assess how well our proposed weighted IPCW estimator performs in the real world, we selected a variable from this data set which was measured on almost the entire cohort and then made it missing for the cohort members not included in the subsample. The variable we chose was Method of Detection. Method of Detection is an indicator of whether the patient or her doctor found her DCIS lesion on physical examination (Palpation) *ver-*

Table 2: *Simulation Study 3.2.* The cohort's absolute risk of having an event by time t for $E^* = 0$ and $E^* = 1$ based on a matched case-control study over 2 censoring levels (column 1), 3 sample sizes (column 2), and 3 time points (columns 4-9). Estimates for the IPCW method (Std. error in parens), the truth, and MSE are based on an average of 100 bootstrap samples.

			5 years		10 years		15 years	
Cens	n	Method	$E* = 0$	$E* = 1$	$E* = 0$	$E* = 1$	$E* = 0$	$E* = 1$
70%	500	IPCW	0.163(.03)	0.322(.08)	0.239(.05)	0.567(.12)	0.311(.08)	0.681(.14)
		Truth	0.167	0.3	0.243	0.535	0.343	0.635
		MSE	.006	.001	.015	.003	.023	.007
	1000	IPCW	0.14(.02)	0.282(.04)	0.261(.03)	0.505(.07)	0.288(.04)	0.647(.09)
		Truth	0.144	0.261	0.264	0.45	0.339	0.597
		MSE	.002	0	.008	.001	.01	.004
	5000	IPCW	0.133(.01)	0.277(.02)	0.246(.01)	0.474(.03)	0.342(.02)	0.603(.04)
		Truth	0.129	.292	0.239	0.493	0.334	0.631
		MSE	0	0	.001	0	.002	0
80%	500	IPCW	0.166(.04)	0.308(.08)	0.225(.04)	0.617(.10)	0.344(.07)	0.836(.11)
		Truth	0.167	0.3	0.243	0.535	0.343	0.635
		MSE	.007	.001	.018	.002	.052	.004
	1000	IPCW	0.15(.02)	0.26(.05)	0.348(.05)	0.446(.08)	0.373(.05)	0.616(.11)
		Truth	0.144	0.261	0.264	0.45	0.339	0.597
		MSE	.002	.001	.006	.009	.012	.004
	5000	IPCW	0.133(.01)	0.292(.02)	0.243(.01)	0.501(.04)	0.301(.02)	0.647(.04)
		Truth	0.129	0.292	0.239	0.493	0.334	0.631
		MSE	0	0	.001	0	.002	.002

Table 3: *Absolute Risk Estimation 4.* Estimates based on the entire cohort, the IPCW and an unmatched case-control with all cases (Sampling 1), the IPCW and an unmatched case-control with a random sample of cases (Sampling 2), and the IPCW and a matched case-control (Sampling 3), for Method of Detection (column 2), number of cases (column 3), number of controls (column 4), odds ratio for Palpation vs. Mammography (column 5), and 5-year absolute risk of recurrence (column 6).

Sample	Detect	Cases	Controls	OR	5 years
Cohort	Palpation	42	126	1.05	0.26
	Mammography	168	527		0.15
IPCW	Palpation	42	68	0.99	0.26
Sampling 1	Mammography	168	269		0.15
IPCW	Palpation	30	62	1.06	0.27
Sampling 2	Mammography	120	262		0.16
IPCW	Palpation	41	54	1.22	0.31
Sampling 3	Mammography	160	258		0.19

sus mammography examination (Mammography). To investigate potential differences between case-control designs, we looked at three different ways of choosing the case-control members: *Sampling* 1 denotes an unmatched design with all cases and a random sampling of controls as described in Section 2.2 and similar to a case-cohort design; *Sampling* 2 denotes an unmatched design with a random sampling of both cases and controls; and *Sampling* 3 denotes a matched design with the matching based on diagnosis year. These three ways of sampling were compared to the 'true' absolute risk of the cohort, i.e., the absolute risk based on *all* measurements of Method of Detection. The results for each of the samplings and the 'true' absolute risk are shown in Table 3.

The results shown in column 5 of Table 3 do not correspond with the odds ratios for Method of Detection reported in Table 1 of Kerlikowske et al. (Kerlikowske et al., 2003) because the latter were adjusted for age at diagnosis which is not the case here. However, the absolute risk of recurrence at 5 years as reported here for Sampling 1 and 2 do fall into the confidence intervals of the absolute risk estimate reported in Table 4 of Kerlikowske et al. (2003) (i.e., in the cited article the 95% confidence interval for Palpation is

$(0.145, 0.275)$ and for Mammography it is $(0.139, 0.196)$).

As seen in Table 3, the cohort's absolute risk of recurrence at 5 years is 0.26 for those who found a mass by palpation and 0.15 for those who found a mass by mammography. In Sampling 1, an unmatched case-control design where all recurrences from the cohort are included as cases and a random sampling of non-recurrences are included as controls, there were almost two controls for each case. The IPCW estimator performed perfectly in this design compared to the entire cohort (i.e., absolute risk of 0.26 for palpation and 0.15 for mammography) even though the odds ratio decreased from 1.05 (in the cohort) to 0.99 (in this unmatched design). The results for Sampling 1 can also be seen in Figure 4. Sampling 2, an unmatched case-control design where both the recurrences and non-recurrences were randomly sampled has approximately one case for every two controls. The IPCW estimator did quite well in this sampling (i.e., absolute risk of 0.27 for palpation and 0.16 for mammography). The odds ratio was almost the same for this sampling as in the cohort (1.06 and 1.05, respectively).

Sampling 3's absolute risk estimate does a bit worse than the cohort's (0.31 and 0.19 *versus* 0.26 and 0.15, respectively). There are two possible reasons for this discrepancy. First, the number of Palpation controls included in Sampling 3 is substantially smaller than the number included in the other two samplings and the cohort (see column 4 in Table 3). As opposed to a one-case-to-two-controls design, Sampling 3 represents more of a one-case-to-one-control design. Second, due to the fewer number of members in the case-control the data is sparser and this may violate the $Pr(\bar{G}_0(T|W) > 0.1) = 1$, a condition for the IPCW method (Section 2.4).

To investigate the first possible cause for discrepancy (and potentially the second), we randomly assigned a few more non-recurrences as controls to increase the total number of controls included. The results are shown in Table 4. There are a total of 11 controls added to the Palpation category as well as one case. As expected, this increase in controls leads to a decrease in the odds ratio from 1.22 to 1.01. Although the 5-year absolute risk estimate does not change for Mammography it does decrease from 0.31 to 0.26 for Palpation. The prevailing discrepancy for Mammography can be seen in Figure 5. Nevertheless, this estimate still falls within the confidence interval of Table 4 of Kerlikowske et al. (2003) (i.e., $(0.139, 0.196)$).

A very important aspect of Sampling 3's results is that the odds ratio of 1.22 (Table 3) is quite biased as an estimate for the cohort's odds ratio of 1.05; however, the absolute risk estimate is not that far from that of

Table 4: *Sampling 3 with a different sampling scheme for the case-control (Section 4).* Estimates based on the IPCW and a matched case-control (Sampling 3) for Method of Detection (column 1), number of cases (column 2), number of controls (column 3), odds ratio for Palpation vs. Mammography (column 4), and 5-year absolute risk of recurrence (column 5).

Detect	Cases	Controls	OR	5 years
Palpation	42	65	1.01	0.26
Mammography	158	246		0.19

the cohort's. This reinforces the use of the proposed method which is not as reliant on disease prevalence (i.e., as measured by the relative risk) as previously documented methods.

Increasing the number of controls seems to have alleviated most of the discrepancy seen in Table 3. An interesting insight given by this data analysis is that the initial odds ratio estimate for Sampling 3 (Table 3) may point to a biased sampling for those chosen in the case-control subsample. It appears that this biased sampling was remedied by increasing the number of controls, i.e., it went from about a one-to-one to a almost a one-to-two sampling scheme.

5 Discussion & Conclusions

We introduced an estimator of the cohort's absolute risk of an event by time t based on variables of interest only measured in a subsample of the cohort. In addition to this method being resilient to prevalence of disease (by not inheriting the relative risk estimate), it allows for informative censoring, is a locally efficient estimator, and collapses to the full data estimator in the presence of the *full data*.

This estimator allows the numerator and denominator to be evaluated separately, resulting in a ratio of two estimators; therefore, a weighting may be necessary to guarantee the probability of no event is equal to 1 at time 0. This can be achieved by dividing the ratio in Equation 11 by the parameter of interest evaluated at 0. Interestingly, a ramification of this weighting is that the denominator of Equation 11, i.e., $\mu(\delta) = Pr(E^* = \delta)$, divides out.

This means that any misspecification of $Pr(E^* = \delta)$ is irrelevant as well as the matching (or non-matching) scheme. Thus, as shown in Section 2.4, the proposed weighted IPCW estimator (Equation 12) can be implemented in either a matched or unmatched case-control subsample design.

To explore the validity of $\vartheta_w(t, \delta)$, we implemented simulation studies over numerous sample sizes and several censoring percentages. The simulation studies enable us to compare $\vartheta_w(t, \delta)$'s estimated absolute risk to the truth. In Tables 2 and 1 as well as Figures 2 and 3 one can see that the IPCW estimator is doing quite well approximating the truth. This is true for finite sample sizes (e.g., $n = 500$) as well as relatively large sample sizes (e.g., $n = 5000$) (Molinaro, 2004). Importantly the latter indicates that although we assume i. i. d. observations in the presence of matching (i.e., there is a fixed ratio between the number of cases and controls), the proposed method is asymptotically valid.

In Section 4, the proposed weighted IPCW estimator was evaluated in the DCIS study. The variable, Method of Detection, was selected as the risk factor of interest. Because this variable was measured in almost all members of the cohort, we could assess our proposed estimators accuracy based on only the measurements for patients in the case-control compared to the those for the entire (or almost entire) cohort. As shown in Tables 3, for Samplings 1 and 2 the proposed estimator is doing well. The discrepancy seen in the estimate for Sampling 3 could be indicative of a biased underlying sampling of the subsample or simply sparsity in the data as there are fewer included controls. To examine this situation, more non-recurrences were added to the sampling. The results in Table 4 show that in fact this did close the gap between the cohort's known absolute risk and odds ratio and that of Sampling 3's.

Thus, limitations of this method will be dependent on the underlying sampling of the subsample. As we saw in Section 4, if there is an unmatched case-control with all the cases or a sampling of the cases the estimation of the full cohort's absolute risk was accurate. However, sparsity in the data, due to matching or to too few participants in the case-control, may violate assumptions needed for the IPCW estimator to be consistent.

In future simulations, the effect of the underlying sampling for the nested case-control, case-cohort, and two-stage case-control study designs will be studied. Additionally, we will implement the locally efficient estimator for continuous and time-dependent variables, as well as expand the proposed method to encompass more than one outcome or disease of interest, i.e.,

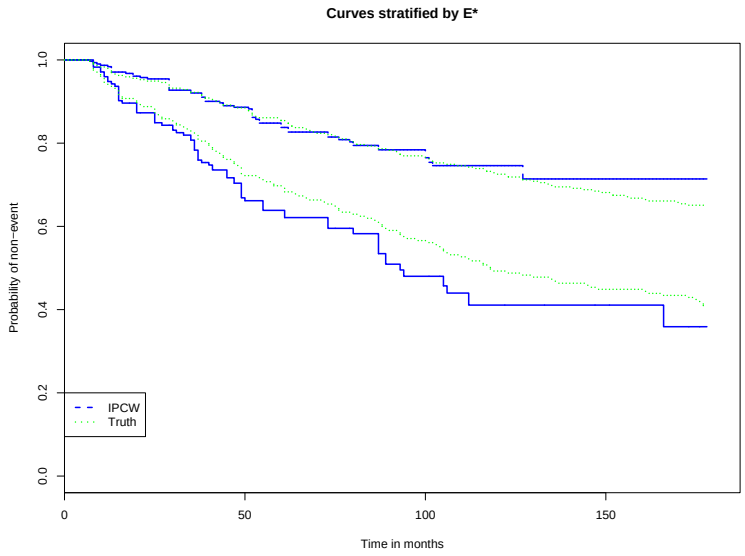

Figure 2: *Survival Curves for Unmatched Case Control (n=500).* The survival curves were calculated for the truth (no cens) and the proposed IPCW method using Δ^{global}, 70% censoring, and cohort size = 500. The top two lines correspond with $E^* = 0$ while the bottom two correspond with $E^* = 1$.

competing risks.

[0] The authors would like to thank Mitchell Gail, Ruth Pfeiffer, and Sholom Wacholder for fruitful discussions. This work was supported by in part by an NIH grant (R01 GM67233), the NCI-funded UCSF Breast Cancer SPORE (P50 CA58207) and Technical support was received from the UCSF Cancer Center (P30 CA82103).

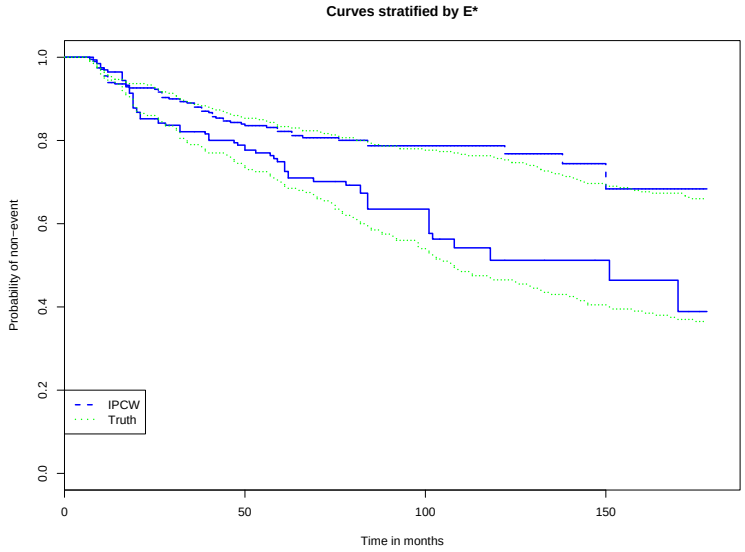

Figure 3: *Survival Curves for Matched Case Control (n=500).* The survival curves were calculated for the truth (no cens) and the proposed IPCW method using Δ^{global}, 70% censoring, and cohort size = 500. The top two lines correspond with $E^* = 0$ while the bottom two correspond with $E^* = 1$.

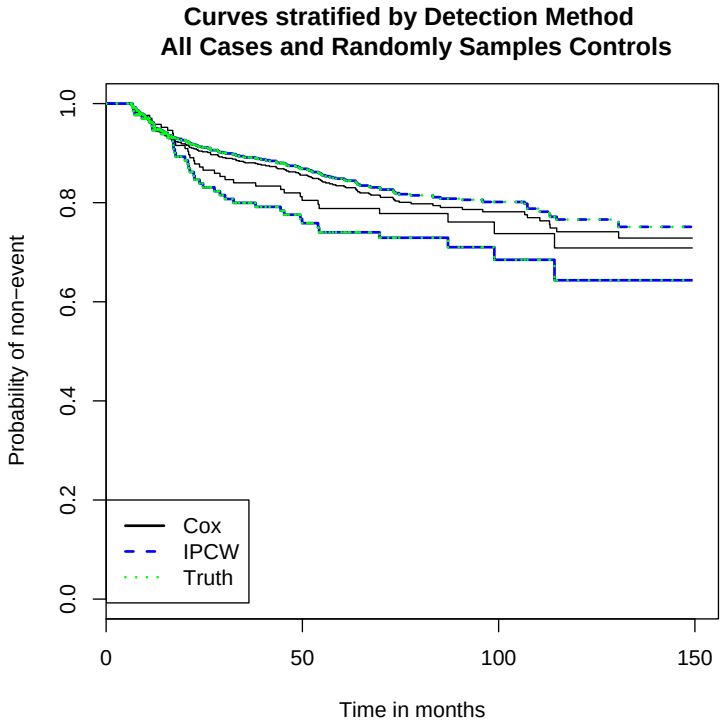

Figure 4: *Survival Curves for Sampling 1.* The survival curves were calculated for the entire cohort (Truth) and the proposed IPCW method stratified by Detection Method using all cases and a random selection of controls. The top two lines correspond with Mammography while the bottom two correspond with Palpation.

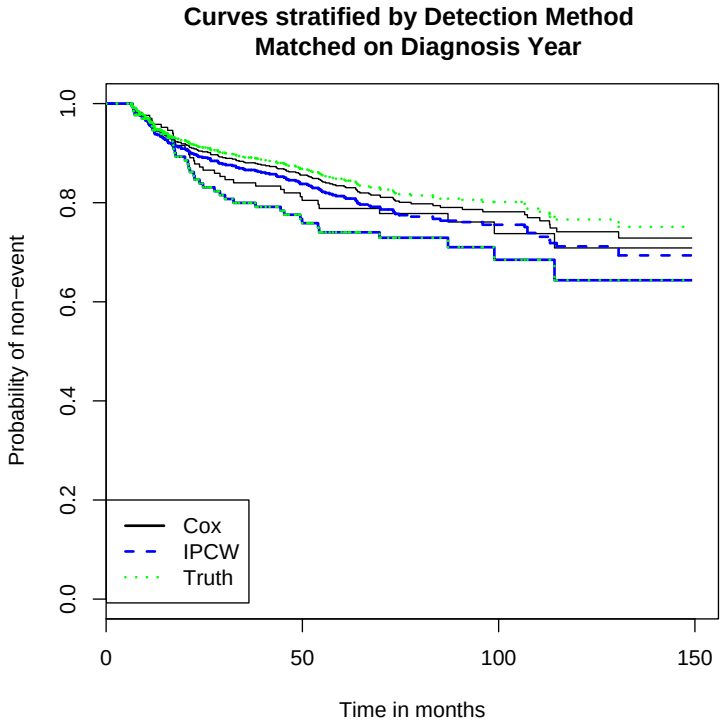

Figure 5: *Survival Curves for Sampling 3 (Additional Controls).* The survival curves were calculated for the entire cohort (Truth) and the proposed IPCW method stratified by Detection Method using a matched design based on year of diagnosis with additional controls added. The top two lines correspond with Mammography while the bottom two correspond with Palpation.

References

Benichou J, Gail MH. Estimates of Absolute Cause-Specific Risk in Cohort Studies. *Biometrics* 1990; **46**: 813–826.

Benichou J, Gail MH. Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics* 1995; **51**: 182–194.

Benichou J, Byrne C, Gail MH. An approach to estimating exposure-specific rates of breast cancer from a two-stage case-control study within a cohort. *Statistics in Medicine* 1997; textbf16: 133–151.

Borgan O, Langholz B. Nonparametric Estimation of Relative Mortality From Nested Case-Control Studies. *Biometrics* 1993; **49**: 593–602.

Borgan O, Goldstein L, Langholz B. Methods for the Analysis of Sampled Cohort Data in the Cox Proportional Hazards Model. *The Annals of Statistics* 1995; **23**: 1749–1778.

Breslow NE, Cain KC. Logistic Regression for two-stage case-control data. *Biometrika* 1988; textbf75(1): 11–20.

Breslow NE, Zhao LP. Logistic Regression for stratified case-control studies. *Biometrics* 1988; textbf44: 891–899.

Breslow NE. Statistics in Epidemiology: The Case-Control Study. *Journal of the American Statistical Association* 1996; **91**(433): 14–28.

Ernster VL. Nested case-control studies. *Preventive Medicine* 1994; **23**: 587–590.

Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* 1991; textbf10: 891–899.

Gill RD, van der Laan MJ, Robins JR. Coarsening at Random: Characterizations, Conjectures and Counter-Examples. In *Proceedings of the First Seattle Symposium in Biostatistics,* Lin DY, Fleming TR (eds). Springer Lecture Notes in Statistics: 1997; 255–294.

Goldstein L, Langholz B. Asymptotic Theory for Nested Case-Control Sampling in the Cox Regression Model. *The Annals of Statistics* 1992; **20**: 1903–1928.

Kerlikowske K, Molinaro AM, Cha I, Ljung BM, Ernster V, Stewart K, Chew K, Moore DH, Waldman F. Predictors of Recurrence Among Women with DCIS Treated by Lumpectomy. *Journal of National Cancer Institute* 2003; **95** (22): 1692–1702.

Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. *Statistical Science* 1996; **11**(1): 35–53.

Langholz B, Borgan O. Estimation of absolute risk from nested case-control data. *Biometrics* 1997; **53**: 767–774.

Liddel JR, McDonald JC, Thomas DC. Methods of Cohort Analysis: Appraisal by Application to Asbestos Mining. *Journal of the Royal Statistical Society, Series A* 1977; **140**: 469–491.

Mantel N. Synthetic Retrospective Studies and Related Topics. *Biometrics* 1973; **29**: 479–486.

Molinaro AM. Novel approaches to prediction of survival in cancer research: Focus on genomics. PhD Thesis, University of California, Berkeley 2004.

Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**: 1–11.

Robins JR, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology, Methodological issues* Bikhauser: 1992.

Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics* 1988; **16**(1): 64–81.

van der Laan MJ, Robins JR. *Unified Methods for Censored Longitudinal Data and Causality.* Springer, 2002.

Wacholder S. Practical consideratons in choosing between the case-cohort and nested case-control design. *Epidemiology* 1991; **2**: 155–158.

Wacholder S, Weinberg CR. Flexible Maximum Likelihood Methods for Assessing Joint Effect in Case-Control Studies with Complex Sampling. *Biometrics* 1994; **50**: 350–357.

Zhao LP, Lipsitz S. Design and analysis of two-stage studies. *Statistics in Medicine* 1992; textbf11: 769–782.