Evaluating prognostic accuracy of biomarkers in nested case–control studies

TIANXI CAI[†]

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

YINGYE ZHENG[*,†]

Department of Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
yzheng@fhcrc.org

SUMMARY

Nested case–control (NCC) design is used frequently in epidemiological studies as a cost-effective sub-cohort sampling strategy to conduct biomarker research. Sampling strategy, on the other hand, creates challenges for data analysis because of outcome-dependent missingness in biomarker measurements. In this paper, we propose inverse probability weighted (IPW) methods for making inference about the prognostic accuracy of a novel biomarker for predicting future events with data from NCC studies. The consistency and asymptotic normality of these estimators are derived using the empirical process theory and convergence theorems for sequences of weakly dependent random variables. Simulation and analysis using Framingham Offspring Study data suggest that the proposed methods perform well in finite samples.

Keywords: Inverse probability weighting; Nested case–control study; Time-dependent accuracy.

1. INTRODUCTION

Novel markers promise to dramatically change the decision-making process in disease monitoring and treatment selection. Prospective cohort studies are crucial to establishing the value of a biomarker for predicting future events such as disease onset and recurrence, or patient survival. To enable future biomarker research, the biospecimens of the full cohort are often collected at baseline and stored for future studies. Many well-known cohort studies adopted such strategies. Examples include the Women's Health Initiative Study (Johnson *and others*, 1999), the Nurses' Health Initiative Study, and the Framingham Offspring Study (Kannel *and others*, 1979). However, since the assessment of biomarkers can be expensive and labor intensive, a standard full cohort design may be infeasible or inefficient for subsequent biomarker studies. To overcome such difficulties, the nested case–control (NCC) study design (Thomas, 1977; Prentice and Breslow, 1978) is often adopted as a cost-effective cohort sampling strategy. Under such a study design, the biospecimen is assayed for all study cases but only for a fraction of controls selected randomly from the risk set of the matched cases.

---

[*]To whom correspondence should be addressed.
[†]These authors contributed equally to this work.

To analyze data from NCC studies, conditional logistic regression is typically used to estimate relative risk parameters, which are equivalently hazard ratio parameters under the proportional hazards (PH) model. Stratified sampling of controls in NCC designs has been considered (Langholz and Borgan, 1995) in order to improve efficiency of a simple NCC design. For the class of estimators based on the partial likelihood under NCC designs, asymptotic properties have been formally derived for estimators of both hazard ratios (Goldstein and Langholz, 1992) and absolute risks (Langholz and Borgan, 1997), using counting process and martingale theory (Andersen and Gill, 1982).

An alternative approach is to consider the class of inverse probability weighted (IPW) estimators, where contributions of individuals are weighted inversely proportional to their sampling fractions. Compared with the partial likelihood–based method, IPW estimators can be more efficient as more individuals are included in risk sets. They are also more flexible in estimating functions beyond hazard ratios. Samuelsen (1997) proposed IPW partial likelihood estimators with weights accounting for NCC sampling; however, theoretical justification has not been formally developed. Recently, much progress has been made in the development of asymptotic theory for IPW estimator when fitting the Cox regression model using survey data obtained using two-phase sampling (Lin, 2000; Breslow and Wellner, 2007). In particular, Breslow and Wellner (2007) developed a modern theory on the conditions required for weak convergence of IPW empirical process under Cox regression with stratified case–cohort sampling. While it could be conjectured that the general theory developed there may be extended to other more complex sampling designs such as the stratified NCC sampling designs considered in this manuscript, to our knowledge, there does not exist a general theory that can be directly applied to IPW process under NCC settings.

Furthermore, most of the current developments in NCC study literature focus primarily on estimating relative risk parameters, with little attention given to other summaries of the data. Prognostic biomarker studies have recognized that the relative risk from risk modeling does not fully assess biomarker performance (Pepe *and others*, 2004). Two classes of time-dependent accuracy measures, retrospective and prospective accuracies, are commonly calculated in these settings. Specifically, for a putative marker Y measured at baseline, its retrospective accuracies, such as the true-positive fraction (TPF) and false-positive fraction (FPF), evaluated at a future predicting time t and a cutpoint c, are defined as

$$\text{TPF}_t(c) = P\{Y > c | T \leqslant t\} \quad \text{and} \quad \text{FPF}_t(c) = P\{Y > c | T > t\}.$$

A time-dependent receiver operating characteristic (ROC) curve, $\text{ROC}_t(u) = \text{TPF}_t\{\text{FPF}_t^{-1}(u)\}$ for $u \in (0, 1)$, is a plot of $\text{TPF}_t(c)$ versus $\text{FPF}_t(c)$ for all c. The prospective accuracy summaries, positive predictive value (PPV) and negative predictive value (NPV), are time-dependent functions of the form

$$\text{PPV}_t(c) = P\{T \leqslant t | Y > c\} \quad \text{and} \quad \text{NPV}_t(c) = P\{T > t | Y \leqslant c\}. \tag{1.1}$$

The prospective accuracy summaries are of more interest to the end users of the test since they quantify the subject's risk of an outcome by time t, given a positive or a negative test result. Statistical models for estimating these quantities using time-to-event data from full cohorts have been proposed (Heagerty *and others*, 2000; Heagerty and Zheng, 2005; Moskowitz and Pepe, 2004; Zheng *and others*, 2008). However, those current methods do not consider subcohort sampling schemes and have only been developed for standard settings with biomarker values fully observed for individuals in the cohort. To efficiently evaluate novel biomarkers for risk prediction, there is an urgent need to expand the research to accommodate various subcohort sampling schemes.

The primary aim of this paper is to formulate and lay foundations for biomarker accuracy estimation with data from NCC studies. We develop estimators for time-dependent prediction accuracy measures based on the idea of inverse probability weighting (Robins *and others*, 1994; Samuelsen, 1997). The asymptotic properties of these estimators are formally derived using convergence theorems for sequences

of negatively associated (NegA) dependent random variables (Liang *and others*, 2004; Liang and Baek, 2006). The remainder of the paper is organized as follows: we present the estimation procedures under both a simple NCC design and a stratified NCC design in Section 2; asymptotic theory is given in Section 3; Section 4 contains numerical studies; an application to the Framingham Offspring Study is given in Section 5, and we conclude with some remarks in Section 6.

2. THE MODEL AND ESTIMATION

2.1 *General notation*

Suppose we have a cohort of N individuals followed prospectively for a clinical event of interest. Due to censoring, the observed event time data consist of N i.i.d. bivariate vector $\{(X_i, \delta_i), i = 1, \ldots, N\}$, where $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leqslant C_i)$, and T_i and C_i denote the event time and censoring time, respectively. Under a standard NCC design, all individuals observed to have an event are selected as "cases" for further evaluation of marker Y with their event times denoted as $\{t_1, \ldots, t_n\}$. In addition, at each selected case's failure time t_j, a random sample of size m is selected without replacement from the risk set $R(t_j) = \{i : X_i \geqslant t_j\}$ as potential controls for marker measurement. The number of individuals at risk at the selected event time t_j is denoted $\mathfrak{n}(t_j)$, with $\mathfrak{n}(t_j) = \sum_i^N I(X_i \geqslant t_j)$. Furthermore, let V_i be a binary random variable with $V_i = 1$ if subject i is ever sampled into the NCC subcohort either as a case or as a control.

2.2 *Estimation with a cohort study*

Using Bayes' theorem, we can rewrite the aforementioned retrospective prognostic accuracy summaries as

$$\mathrm{TPF}_t(c) = \frac{\{1 - \mathcal{F}(c)\} - \mathcal{S}(t, c)}{1 - \mathcal{S}(t)} \quad \text{and} \quad \mathrm{FPF}_t(c) = \frac{\mathcal{S}(t, c)}{\mathcal{S}(t)}$$

and the prospective accuracy summaries as

$$\mathrm{PPV}_t(c) = \frac{\{1 - \mathcal{F}(c)\} - \mathcal{S}(t, c)}{1 - \mathcal{F}(c)} \quad \text{and} \quad \mathrm{NPV}_t(c) = \frac{\{\mathcal{S}(t) - \mathcal{F}(c)\} - \mathcal{S}(t, c)}{\mathcal{F}(c)},$$

where $\mathcal{S}(t, c) = P(T > t, Y > c)$ is the bivariate survival function of T and Y, $\mathcal{F}(c) = P(Y \leqslant c)$ is the marginal cumulative distributional function of marker Y, and $\mathcal{S}(t) = P(T > t) \equiv \mathcal{S}(t, -\infty)$ is the marginal survival function of T. Therefore, the key components for calculating all 4 accuracy summaries involve $\mathcal{F}(c)$, as well as the bivariate survival function $\mathcal{S}(t, c)$ for specific values of c and t.

When Y was measured for all individuals in the cohort, plug-in estimators of the accuracy measures had been proposed using either nonparametric or semiparametric estimators of these functionals (Heagerty *and others*, 2000; Zheng *and others*, 2008, 2010). Specifically, one can estimate the marginal distribution of Y empirically as

$$\widehat{\mathcal{F}}(y) = \frac{1}{N} \sum_{i=1}^{N} I(Y_i > y) \tag{2.2}$$

and the bivariate survival function $\mathcal{S}(t, c)$ as

$$\widehat{\mathcal{S}}(t, c) = \int_c^\infty \widehat{\mathcal{S}}(t|y) \mathrm{d}\widehat{\mathcal{F}}(y) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{S}(t|Y_i) I(Y_i > c), \tag{2.3}$$

where $\widehat{S}(t|y)$ is an estimator for the conditional survival function $S(t|y) = P(T > t|Y = y)$, relating to the absolute risk by time t given y. Under a PH model for survival time T and marker Y: $\lambda(t) = \lambda_0(t)\exp(\beta Y)$, $S(t|y)$ can be estimated as $\widehat{S}(t|y) = \exp\{-\widehat{\Lambda}_0(t)\exp(\widehat{\beta}y)\}$, where $\widehat{\beta}$ is estimated from the partial likelihood and

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^{N} \frac{\delta_i I(X_i \leqslant t)}{\sum_{j=1}^{N} I(X_j \geqslant X_i)\exp(\widehat{\beta}Y_j)}$$

is the Breslow estimator of $\Lambda_0(t)$. Plug-in estimators $\widehat{\text{TPR}}$, $\widehat{\text{FPR}}$, $\widehat{\text{PPV}}$, and $\widehat{\text{NPV}}$ can be obtained based on (2.2) and (2.3).

2.3   *Estimation under an NCC design*

Under an NCC design, since Y is selectively measured depending on outcome and other covariates, rather than completely at random, it is crucial to adjust for the sampling scheme of NCC in order to provide unbiased estimates of the aforementioned accuracy summaries. We will achieve this by casting the problem into the general framework of a failure-time regression model with missing covariates and develop IPW estimators for $S(t, c)$ and $\mathcal{F}(c)$. Our IPW-based procedures will provide consistent estimators of $S(t, c)$ and $\mathcal{F}(c)$ and consequently valid estimators for time-dependent accuracy summaries.

Sampling probability of an NCC design.   As a simple case of the IPW procedure, we consider weighing the contributions from the selected observations with weight $\widehat{w}_i = V_i/\widehat{p}_i$, where $\widehat{p}_i$ is the probability of the ith subject being selected to the NCC cohort based on the sampling scheme. When all individuals in the cohort with $\delta = 1$ are included in NCC samples, the selection probability is $\widehat{p}_i = \delta_i + (1 - \delta_i)\{1 - \widehat{G}(X_i)\}$ as given by Samuelsen (1997), where

$$\widehat{G}(X_i) = \prod_{j:X_j < X_i} \left\{1 - \frac{m\delta_j}{\mathfrak{n}(X_j) - 1}\right\}.$$

It is easy to show that $E\{V_i/\widehat{p}_i|(X_i, \delta_i)\} = 1$. Note that this is the "true" weights dictated by the NCC design. Other weights can be considered. For example, Chen (2001) suggested using local averaging weights within intervals of censoring times. Such "estimated" weight has been shown to improve efficiency with simulations especially in situations when censoring is dependent on covariates.

Estimating $S(t|y)$.   The log hazard ratio, β, under the PH model can be obtained by maximizing a weighted partial likelihood with weights accounting for outcome-dependent sampling as described above. Specifically, one may estimate β as $\widehat{\beta}^{\text{w}} = \text{argmax}_\beta \mathcal{L}(\beta)$, where

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} \widehat{w}_i \delta_i \left\{\beta Y_i - \log \sum_{j=1}^{N} \widehat{w}_j I(X_j \geqslant X_i)\exp(\beta Y_j)\right\}. \tag{2.4}$$

Based on $\widehat{\beta}^{\text{w}}$, $S(t|y)$ can be estimated as $\widehat{S}^{\text{w}}(t|y) = \exp\{-\widehat{\Lambda}_0^{\text{w}}(t)\exp(\widehat{\beta}^{\text{w}}y)\}$, where

$$\widehat{\Lambda}_0^{\text{w}}(t) = \sum_{i=1}^{N} \frac{\widehat{w}_i I(X_i \leqslant t)\delta_i}{\sum_{j=1}^{N} I(X_j \geqslant X_i)\widehat{w}_j \exp(\widehat{\beta}Y_j)}$$

can be viewed as a weighted Breslow estimator of $\Lambda_0(t)$.

Note that $\widehat{\mathcal{S}}^{w}(t|y)$ is different from an existing absolute risk estimator for NCC design defined as $\widehat{\mathcal{S}}^{LB}(t|y) = \exp\{-\widehat{\Lambda}_0^{LB}(t)\exp(\widetilde{\beta}y)\}$, proposed by Langholz and Borgan (1997), where $\widetilde{\beta}$ is the conditional logistic regression estimator of β,

$$\widehat{\Lambda}_0^{LB}(t) = \sum_{i=1}^{n} \frac{I(X_i \leqslant t)\delta_i}{\sum_{l \in \tilde{R}(t_j)} \exp(\widetilde{\beta}Y_l)\mathfrak{n}(t_j)/(m+1)},$$

and $\tilde{R}(t_j)$ indexes the case who failed at time t_j and the corresponding m matched controls. Intuitively, $\widetilde{\beta}$ and $\widehat{\mathcal{S}}^{LB}(t|y)$ obtained from the above procedures might be less efficient since only the case and m selected controls, rather than all the samples at risk, are used in the partial likelihood and for calculating $\widehat{\Lambda}_0^{LB}(t)$.

Estimating $F_Y(y)$ and $\mathcal{S}(t, c)$. With NCC sampling, we again construct IPW estimators for the distribution function of marker Y and the bivariate survival function of Y and T. Specifically, we consider the following empirical estimator:

$$\widehat{\mathcal{S}}^{w}(t, c) = \frac{\sum_i^N \widehat{\mathcal{S}}^{w}(t|Y_i)\widehat{w}_i I(Y_i > c)}{\sum_{j=1}^N \widehat{w}_i}$$

for $\mathcal{S}(t, c)$, similar to the representation considered in Akritas (1994). Subsequently, we may estimate $\mathcal{F}_Y(c)$ as $\widehat{\mathcal{F}}_Y^{w}(c) = \widehat{\mathcal{S}}^{w}(0, c)$ and the marginal survival distribution of T, $\mathcal{S}(t)$, as $\widehat{\mathcal{S}}^{w}(t) = \widehat{\mathcal{S}}^{w}(t, -\infty)$.

Estimating accuracy summaries. Let $\widehat{\mathcal{S}}^{w}(c, t)$, $\widehat{\mathcal{S}}^{w}(0t)$, and $\widehat{F}^{w}(c)$ denote the respective estimators of $\mathcal{S}(c, t)$, $\mathcal{S}(t)$, and $F(c)$, plug-in accuracy estimators under an NCC study for retrospective accuracy summaries can be calculated as

$$\widehat{TPF}_t^{w}(c) = \frac{\{1 - \widehat{\mathcal{F}}^{w}(c)\} - \widehat{\mathcal{S}}^{w}(t, c)}{1 - \widehat{\mathcal{S}}^{w}(t)} \quad \text{and} \quad \widehat{FPF}_t^{w}(c) = \frac{\widehat{\mathcal{S}}^{w}(t, c)}{\widehat{\mathcal{S}}^{w}(t)}$$

and $ROC_t^{w}(u) = \widehat{TPF}_t^{w}\{\widehat{FPF}_t^{w-1}(u)\}$. The estimators for prospective accuracy summaries are

$$\widehat{PPV}_t^{w}(c) = \frac{\{1 - \widehat{\mathcal{F}}^{w}(c)\} - \widehat{\mathcal{S}}^{w}(t, c)}{1 - \widehat{\mathcal{F}}^{w}(c)} \quad \text{and} \quad \widehat{NPV}_t^{w}(c) = \frac{\{\widehat{\mathcal{S}}^{w}(t) - \widehat{\mathcal{F}}^{w}(c)\} - \widehat{\mathcal{S}}^{w}(t, c)}{\widehat{\mathcal{F}}^{w}(c)}.$$

Note that an alternative strategy to estimate the accuracy summaries is to replace $\widehat{\mathcal{S}}^{w}(t|Y_i)$ in $\widehat{\mathcal{S}}^{w}(t, c)$ with $\widehat{\mathcal{S}}^{LB}(t|Y_i)$. IPW is still needed here in order to retrieve information on the marginal distribution of Y.

2.4 *Stratified NCC sampling*

Often in practice some covariate information is available for all cohort members. A stratified sampling of controls based on surrogate variables that are correlated with the marker may enhance the power of a simple NCC design. In a stratified NCC sampling, at each selected case's failure time, controls are selected for marker measurement randomly without replacement among those who are in the risk set and matched to the case based on some covariate Z. Without loss of generality, we assume that Z consists of discrete random variables. To incorporate additional matching in the proposed IPW approach, we replace $\widehat{w}_i$ in the weighted likelihood (2.4) with $\widehat{w}_{zi} = V_i/\widehat{p}_{zi}$, where $\widehat{p}_{zi} = \delta_i + (1 - \delta_i)\{1 - \widehat{G}_z(X_i, Z_i)\}$,

$$\widehat{G}_z(X_i, Z_i) = \prod_{j: X_j < X_i, Z_j = Z_i} \left\{1 - \frac{m\delta_j}{\mathfrak{n}_z(X_j, Z_j) - 1}\right\},$$

and $\mathfrak{n}_z(X_j, Z_j) = \sum_{k=1}^{N} I(X_k \geqslant X_j, Z_k = Z_j)$ is the size of the covariate matched risk set for failure time X_j. It is easy to show that $E\{V_i/\widehat{p}_i | (X_i, \delta_i, Z_i)\} = 1$. Note that under a stratified NCC sampling design, estimation procedures for obtaining $\widetilde{\beta}$ and $\widehat{\mathcal{S}}^{\mathrm{LB}}(t|y)$ remain the same, which is clear from equation (4.2) in Borgan *and others* (1995).

3. INFERENCE PROCEDURES

Deriving inference procedures for the proposed IPW-based estimators for accuracy summaries can be challenging due to the complex data structure induced by NCC sampling. In particular, under the finite-population sampling scheme, indicators of being sampled, V_i, are weakly dependent conditional on $\mathscr{D} = \{(X_i, \delta_i, Y_i, Z_i), i = 1, \ldots, N\}$. The standard convergence theorems, such as the law of large numbers or central limit theorems for i.i.d. cases, are not directly applicable here. In the supplementary material available at *Biostatistics* online, we provide more detailed justifications for the consistency and asymptotic normality of proposed IPW estimators for accuracy summaries for NCC design with finite-population sampling, using results on the strong and weak convergence of weighted sums of NegA dependent variables (Liang *and others*, 2004; Liang and Baek, 2006). The key is to show that $\widehat{U} = N^{-1} \sum_{i=1}^{N} \widehat{w}_i R_i$, where $R_i = R(X_i, \delta_i, Y_i)$ for some deterministic function R, can be viewed as weighted sums of NegA dependent variables, and it satisfies the conditions required for tightness and weak convergence of the NegA process (see Appendix B of the supplementary material available at *Biostatistics* online).

To obtain interval estimates of specific components of our proposed IPW estimators for $\mathrm{TPF}_t(c)$, $\mathrm{FPF}_t(c)$, $\mathrm{ROC}_t(u)$, $\mathrm{PPV}_t(c)$, and $\mathrm{NPV}_t(c)$, we show in Appendix A of the supplementary material (available at *Biostatistics* online) that for any of these accuracy measures, denoted by a generic term $\mathcal{A}$, $N^{\frac{1}{2}}(\widehat{\mathcal{A}} - \mathcal{A}) = N^{-\frac{1}{2}} \sum_{i=1}^{N} \widehat{w}_i U_{\mathcal{A}i} + o_p(1)$, which is asymptotically normal with mean 0 and variance

$$\sigma_{\mathcal{A}}^2 = E\left(\frac{U_{\mathcal{A}i}^2}{p_i}\right) - m\mathcal{R}_{U_{\mathcal{A}}}^2 = E(U_{\mathcal{A}i}^2) + E\left\{\sigma_{\widehat{w}_i|\mathscr{D}}^2 U_{\mathcal{A}i}^2\right\} - m\mathcal{R}_{U_{\mathcal{A}}}^2,$$

where p_i is the limiting value of $\widehat{p}_i$, $\sigma_{\widehat{w}_i|\mathscr{D}}^2$ is the conditional variance of $\widehat{w}_i$ given $\mathscr{D}$, and $\mathcal{R}_{U_{\mathcal{A}}}^2$ as a functional of $U_{\mathcal{A}}$ is defined in Appendix A of the supplementary material (available at *Biostatistics* online). The term $E(U_{\mathcal{A}i}^2)$ in $\sigma_{\mathcal{A}}^2$ represents the variance of $\widehat{\mathcal{A}}$ if Y is observed for the full cohort, $E\{\sigma_{\widehat{w}_i|\mathscr{D}}^2 U_{\mathcal{A}i}^2\} - m\mathcal{R}_{U_{\mathcal{A}}}^2$ represents the inflated variance due to the missing information on Y for the weighted estimator, and $-m\mathcal{R}_{U_{\mathcal{A}}}^2$ represents the variance adjustment due to the weak negative correlation among the V_i's. Omitting the last term from $\sigma_{\mathcal{A}}^2$ would lead to the robust variance estimator of $\widehat{\mathcal{A}}$ by treating $\widehat{w}_i$ as known independent weights. Thus, using the robust variance alone would always overestimate the true variance of $\widehat{\mathcal{A}}$.

In Appendix C of the supplementary material (available at *Biostatistics* online), we derive $U_{\mathcal{A}i}$ for $\mathcal{A} = \mathrm{TPR}_t(c)$, $\mathrm{FPR}_t(c)$, $\mathrm{ROC}_t(u)$, $\mathrm{PPV}_t(c)$, and $\mathrm{NPV}_t(c)$, respectively. The asymptotic variance of these accuracy summary estimators can be estimated empirically, and the confidence intervals (CIs) can be constructed based on normal approximations. For example, we show in Appendix C of the supplementary material (available at *Biostatistics* online) that $n^{\frac{1}{2}}\{\widehat{\mathrm{ROC}}_t(u) - \mathrm{ROC}_t(u)\} \to N\{0, \sigma_{\mathrm{ROC}_t}^2(u)\}$ in distribution, where $\sigma_{\mathrm{ROC}_t}^2(u) = E\{U_{\mathrm{ROC}_t i}^2(u)/p_i\} - m\mathcal{R}_{U_{\mathrm{ROC}_t(u)}}^2$. $U_{\mathrm{ROC}_t i}(u)$ and $\mathcal{R}_{U_{\mathrm{ROC}_t(u)}}$ are defined in Appendix C of the supplementary material (available at *Biostatistics* online). A 95% CI for $\mathrm{ROC}_t(u)$ may be obtained as $\widehat{\mathrm{ROC}}_t(u) \pm 1.96 N^{-\frac{1}{2}} \widehat{\sigma}_{\mathrm{ROC}_t}(u)$, where $\widehat{\sigma}_{\mathrm{ROC}_t}^2(u) = N^{-1} \sum_{i=1}^{N} \widehat{w}_i \{\widehat{U}_{\mathrm{ROC}_t}(u; \mathscr{D}_i)^2/\widehat{p}_i\} - m\widehat{\mathcal{R}}_{U_{\mathrm{ROC}_t(u)}}^2$, where $\widehat{U}_{\mathrm{ROC}_t}(u; \mathscr{D}_i)$ and $\widehat{\mathcal{R}}_{U_{\mathrm{ROC}_t(u)}}$ are obtained by replacing all theoretical quantities in $U_{\mathrm{ROC}_t}(u)$ and $\mathcal{R}_{U_{\mathrm{ROC}_t(u)}}$ by their empirical counterparts, respectively.

4. SIMULATIONS

We conduct simulation studies to examine the performance of our proposed methods with practical sample sizes. All results are based on 2000 simulated data sets. We first consider a simple situation with a cohort of $N = 1000$. We simulate Y from a standard normal distribution and generate T from a PH model: $\lambda(t) = 0.1 \exp(\beta Y)$, where $\beta = \log(3)$. Censoring time C is taken to be the minimum of 2 and W, where W follows a gamma distribution, with a shape parameter of 2.5 and a rate parameter of 2. This yields approximately 87% censoring and an average of 133 observed cases. For each cohort, we assemble an NCC subcohort by selecting all cases and m controls per case, with $m = 1$ or 3. We consider estimating $\mathrm{TPF}_t(c)$, $\mathrm{FPF}_t(c)$, $\mathrm{PPV}_t(c)$, and $\mathrm{NPV}_t(c)$ for c equal to the 25th and 75th percentiles of the standard normal distribution and $t = 1$ under such an NCC sampling. We calculate the plug-in estimators for the accuracy measures by estimating $\mathcal{S}(t|y)$ based on either $\widehat{\mathcal{S}}^{\mathrm{LB}}(t|y)$ (referred to as LB estimators) or $\widehat{\mathcal{S}}^{\mathrm{w}}(t|y)$ (referred to as IPW estimators). In addition, analytical naive standard errors, ignoring correlations among V_i, and standard errors adjusted for correlations are calculated for the IPW estimators. The results are presented in Table 1. It appears that both sets of estimators are unbiased. However, the IPW estimators in general are more efficient than the LB estimators, as evidenced by their higher relative efficiencies, defined as the ratio of empirical variance estimated from full cohort data to the specific estimated variance from NCC samples. Our adjusted standard errors performed well, with coverage percentage close to 95%. The naive standard errors in most of the cases are quite close to their adjusted counterparts, indicating

Table 1. *Averages of parameter estimates (Ave), empirical standard error of estimates (SE), the averages of the standard error estimates from naive ($\widehat{SE}^{\mathrm{n}}$) and adjusted ($\widehat{SE}^{\mathrm{a}}$) estimators, their corresponding coverage probabilities (CP^{n} and CP^{a}) using the "IPW" approach, average and empirical standard error using the "LB" approach (Ave_{LB} and SE_{LB}) based on NCC samplings with cohort size of $N = 1000$ and $m = 1$ or 3 controls. Relative efficiencies of both approaches (RE_{ipw} and RE_{LB}) are relative to the cohort. Accuracy summaries of marker Y are evaluated at $t = 1$ and $c_1 = F_Y^{-1}(0.25)$, $c_2 = F_Y^{-1}(0.75)$, respectively. Censoring times are independent of Y*

	True	Ave	SE	$\widehat{SE}^{\mathrm{n}}$	$\widehat{SE}^{\mathrm{a}}$	CP^{n}	CP^{a}	Ave_{LB}	SE_{LB}	RE_{IPW}	RE_{LB}
$m = 1$											
$\widehat{\beta}$	1.10	1.12	14.6	13.4	13.3	0.92	0.92	1.13	19.7	0.48	0.22
$\widehat{\mathrm{TPF}}_t(c_1)$	0.95	0.95	1.46	1.35	1.35	0.93	0.93	0.95	1.68	0.42	0.27
$\widehat{\mathrm{FPF}}_t(c_1)$	0.72	0.71	4.49	4.42	4.40	0.95	0.95	0.71	4.58	0.12	0.11
$\widehat{\mathrm{PPV}}_t(c_1)$	0.18	0.18	1.86	2.12	1.83	0.97	0.94	0.19	2.73	0.71	0.32
$\widehat{\mathrm{NPV}}_t(c_1)$	0.97	0.97	0.73	0.71	0.68	0.95	0.93	0.97	0.79	0.68	0.50
$\widehat{\mathrm{TPF}}_t(c_2)$	0.60	0.60	4.81	4.76	4.75	0.94	0.94	0.60	5.80	0.47	0.31
$\widehat{\mathrm{FPF}}_t(c_2)$	0.19	0.19	3.63	3.51	3.50	0.94	0.93	0.19	3.87	0.14	0.12
$\widehat{\mathrm{PPV}}_t(c_2)$	0.35	0.35	4.40	4.43	4.15	0.96	0.94	0.36	6.99	0.48	0.17
$\widehat{\mathrm{NPV}}_t(c_2)$	0.92	0.92	1.06	1.17	1.04	0.98	0.94	0.92	1.10	0.80	0.71
$m = 3$											
$\widehat{\beta}$	1.10	1.11	10.9	10.8	10.8	0.94	0.94	1.11	13.4	0.70	0.45
$\widehat{\mathrm{TPF}}_t(c_1)$	0.95	0.95	1.04	1.04	1.04	0.96	0.96	0.95	1.18	0.67	0.52
$\widehat{\mathrm{FPF}}_t(c_1)$	0.72	0.71	2.74	2.74	2.74	0.95	0.95	0.71	2.76	0.31	0.30
$\widehat{\mathrm{PPV}}_t(c_1)$	0.18	0.18	1.63	1.72	1.64	0.96	0.95	0.19	1.83	0.85	0.68
$\widehat{\mathrm{NPV}}_t(c_1)$	0.97	0.97	0.59	0.61	0.6	0.96	0.95	0.97	0.65	0.78	0.66
$\widehat{\mathrm{TPF}}_t(c_2)$	0.60	0.60	3.94	3.86	3.85	0.95	0.95	0.60	4.30	0.73	0.59
$\widehat{\mathrm{FPF}}_t(c_2)$	0.19	0.19	2.24	2.22	2.22	0.94	0.94	0.19	2.36	0.35	0.31
$\widehat{\mathrm{PPV}}_t(c_2)$	0.35	0.35	3.37	3.39	3.31	0.95	0.95	0.35	4.25	0.76	0.46
$\widehat{\mathrm{NPV}}_t(c_2)$	0.92	0.92	0.97	1.01	0.97	0.96	0.95	0.92	0.98	0.86	0.84

Table 2. *Summary statistics (same as in Table 1) from simulation studies based on NCC samplings with cohort size of $N = 1000$ and $m = 1$ or 3 controls and with censoring times dependent on Y*

	True	Ave	SE	$\widehat{SE}^n$	$\widehat{SE}^a$	CP^n	CP^a	Ave_{LB}	SE_{LB}	RE_{IPW}	RE_{LB}
$m = 1$											
$\widehat{\beta}$	1.10	1.11	20.0	19.3	19.3	0.94	0.94	1.11	22.7	0.50	0.39
$\widehat{TPF}_t(c_1)$	0.97	0.97	0.93	0.88	0.88	0.93	0.93	0.97	0.94	0.44	0.44
$\widehat{FPF}_t(c_1)$	0.89	0.89	2.27	2.21	2.21	0.93	0.93	0.89	2.30	0.23	0.22
$\widehat{PPV}_t(c_1)$	0.26	0.26	1.78	2.10	1.81	0.98	0.95	0.26	1.81	0.82	0.79
$\widehat{NPV}_t(c_1)$	0.91	0.91	1.85	1.91	1.84	0.95	0.94	0.91	1.97	0.64	0.56
$\widehat{TPF}_t(c_2)$	0.18	0.18	2.61	2.64	2.64	0.95	0.95	0.18	2.91	0.62	0.50
$\widehat{FPF}_t(c_2)$	0.06	0.06	1.49	1.46	1.46	0.93	0.93	0.06	1.50	0.27	0.26
$\widehat{PPV}_t(c_2)$	0.48	0.49	5.75	5.75	5.58	0.94	0.93	0.49	6.68	0.50	0.37
$\widehat{NPV}_t(c_2)$	0.78	0.78	1.63	1.91	1.65	0.98	0.95	0.78	1.58	0.82	0.88
$m = 3$											
$\widehat{\beta}$	1.10	1.10	15.9	15.5	15.5	0.94	0.94	1.11	17.9	0.81	0.64
$\widehat{TPF}_t(c_1)$	0.97	0.97	0.70	0.68	0.68	0.93	0.93	0.97	0.74	0.76	0.68
$\widehat{FPF}_t(c_1)$	0.89	0.89	1.48	1.46	1.46	0.94	0.94	0.89	1.49	0.54	0.54
$\widehat{PPV}_t(c_1)$	0.26	0.26	1.67	1.76	1.70	0.95	0.95	0.26	1.67	0.91	0.91
$\widehat{NPV}_t(c_1)$	0.91	0.91	1.62	1.60	1.58	0.94	0.94	0.91	1.71	0.86	0.77
$\widehat{TPF}_t(c_2)$	0.18	0.18	2.22	2.20	2.20	0.94	0.94	0.18	2.34	0.85	0.77
$\widehat{FPF}_t(c_2)$	0.06	0.06	0.97	0.99	0.99	0.95	0.95	0.06	1.00	0.61	0.58
$\widehat{PPV}_t(c_2)$	0.48	0.48	4.51	4.50	4.46	0.94	0.94	0.48	5.12	0.80	0.62
$\widehat{NPV}_t(c_2)$	0.78	0.78	1.53	1.61	1.56	0.96	0.95	0.78	1.49	0.91	0.96

that correlations are quite weak among observations. However, they do appear to be more conservative when $m = 1$.

In the second scenario (Table 2), T is generated using the same model but Y and C are constructed with a more complicated scheme. Specifically, Y is generated from a mixture of normal distributions: $Y = \mathscr{B}W_1 + (1 - \mathscr{B})W_2$, where $\mathscr{B} \sim$ Bernoulli(0.9), $W_1 \sim N(0, 0.5)$, and $W_2 \sim N(0.3, 0.1)$. To introduce marker-dependent censoring, we let $C \sim$ Uniform(0.5, 1.5) if $\mathscr{B} = 1$ and $C = \exp(\mathscr{Z}/5 - 3Y)$ with $\mathscr{Z} \sim N(0, 1)$ if $\mathscr{B} = 0$. The average number of cases in this setting is around 200. We would expect our IPW estimators to be robust in this situation. Indeed, all estimators are unbiased. The adjusted variance estimators also work well; however, the naive variance estimators for $PPV_t(c)$ and $NPV_t(c)$ again could have inflated values with one matched control.

We also consider a scenario where controls are also matched with cases on a binary covariate Z. With a cohort of size 2000, we first generate Y from a standard normal distribution, we then generate a binary Z such that $Z = 1$ if $Y > 0$. The same models for T and C are used as in the first scenario, which yield on average 266 cases among 2000 simulated data sets. Results in Table 3 again indicate that our proposed estimators under stratified NCC design perform well. The naive variance, without considering the correlation due to finite sampling, often lead to overestimated variances.

5. EXAMPLE

We illustrate our proposal by evaluating the accuracy of an inflammation marker, C-reactive protein (CRP), for predicting the risk of cardiovascular disease (CVD) using the Framingham Offspring Study. Conventional risk factors have been identified for assessing CVD risk in the general population, but these characteristics explain only a fraction of CVD risk. Contemporary biomarkers such as CRP have been sought for use in risk stratification and preventive decision making (Ridker *and others*, 2000). However, its clinical usefulness has not been established.

Table 3. *Summary statistics (same as in Table 1) from simulation studies based on NCC samplings with cohort size of $N = 2000$ and matched on a binary covariate Z*

	True	Ave	SE	$\widehat{SE}^n$	$\widehat{SE}^a$	CP^n	CP^a	Ave_{bl}	SE_{bl}	RE_{ipw}	RE_{bl}
$m = 1$											
β	1.10	1.11	8.60	9.30	8.49	0.96	0.94	1.11	17.07	0.60	0.15
$\widehat{TPF}_t(c_1)$	0.95	0.95	1.23	1.24	1.22	0.93	0.92	0.95	1.56	0.26	0.16
$\widehat{FPF}_t(c_1)$	0.72	0.72	4.79	5.07	4.76	0.95	0.94	0.72	4.93	0.05	0.05
$\widehat{PPV}_t(c_1)$	0.18	0.19	1.47	1.74	1.45	0.98	0.95	0.19	2.75	0.51	0.14
$\widehat{NPV}_t(c_1)$	0.97	0.97	0.50	0.56	0.50	0.96	0.94	0.97	0.64	0.64	0.39
$\widehat{TPF}_t(c_2)$	0.60	0.60	3.23	3.16	3.15	0.94	0.94	0.60	4.51	0.56	0.27
$\widehat{FPF}_t(c_2)$	0.19	0.19	2.00	2.55	2.01	0.99	0.95	0.19	2.48	0.20	0.13
$\widehat{PPV}_t(c_2)$	0.35	0.35	2.56	2.71	2.48	0.97	0.95	0.36	6.29	0.64	0.10
$\widehat{NPV}_t(c_2)$	0.92	0.92	0.73	1.01	0.75	0.99	0.96	0.92	0.82	0.75	0.63
$m = 3$											
β	1.10	1.10	7.10	7.36	7.07	0.95	0.94	1.11	12.15	0.81	0.27
$\widehat{TPF}_t(c_1)$	0.95	0.95	0.90	0.88	0.87	0.93	0.93	0.95	1.17	0.49	0.27
$\widehat{FPF}_t(c_1)$	0.72	0.72	3.06	3.10	2.94	0.95	0.94	0.72	3.14	0.13	0.12
$\widehat{PPV}_t(c_1)$	0.18	0.18	1.23	1.32	1.22	0.97	0.95	0.19	2.04	0.75	0.27
$\widehat{NPV}_t(c_1)$	0.97	0.97	0.43	0.45	0.43	0.96	0.94	0.97	0.51	0.78	0.56
$\widehat{TPF}_t(c_2)$	0.60	0.60	2.60	2.62	2.61	0.95	0.95	0.60	3.45	0.79	0.45
$\widehat{FPF}_t(c_2)$	0.19	0.19	1.33	1.60	1.36	0.98	0.95	0.19	1.63	0.44	0.30
$\widehat{PPV}_t(c_2)$	0.35	0.35	2.15	2.21	2.17	0.95	0.95	0.35	4.52	0.84	0.19
$\widehat{NPV}_t(c_2)$	0.92	0.92	0.69	0.79	0.69	0.97	0.95	0.92	0.75	0.89	0.75

The Framingham Offspring Study was initiated in 1971 with a cohort of 5124 participants. The Framingham Study samples were monitored prospectively, providing a valuable resource for studying epidemiological and genetic risk factors of CVD. We consider here 3289 Offspring Study participants with CRP measurements at the second examination and who were free of CVD at the examination (mean age 44 years and 53% women). We consider outcome as time from examination date to first major CVD event or CVD-related death as defined previously (Lloyd-Jones *and others*, 2004). Since we are interested in the short-term predictive capacity of CRP, individuals with follow-up time longer than 10 years were considered censored at 10 years. During the follow-up period, 251 participants were observed to encounter at least one CVD event. Since CRPs are complete in the cohort, the Framingham data allow us to illustrate our methods with a real data set and to compare the relative efficiency of a few sampling strategies with analysis using the full cohort. From the full cohort data, we further assemble NCC subcohorts with 1 or 3 controls who were selected either (i) without additional matching or (ii) matched to their corresponding cases based on gender and age groups (<30, 30–39, 40–49, >50).

We use a PH model to specify the relation between the failure time and CRP concentration (in log scale). We then calculate the accuracy measures of CRP for predicting CVD by 5 years since the measurement of CRP (i.e., $t = 5$). We considered low and high thresholds, set as the 25th or 75th percentile of the CRP levels in the full cohort. That is, these thresholds would, respectively, classify approximately 75% or 25% of the population as testing positive based on CRP. The accuracy summaries at both values are presented in Table 4, where, in the first and second columns, parameter estimates and standard error estimates for the cohort data are given. The estimates from different NCC samples, presented in the rest of the columns, are quite close to the results using the full cohort data, suggesting that time-dependent accuracy summaries can be reliably estimated from the case–control data with our methods. Matching on age and gender improves the efficiency of most estimates slightly, and including more controls results in more precise inference. Compared to a full cohort analysis, the standard errors for TPF and FPF, although

Table 4. *Estimated accuracy summaries (Ave) and standard errors (SE) of CRP at* 2 *thresholds for predicting* 5-*year CVD events from Framingham data with m* = 1 *and* 3 *controls and with/without stratified by age and gender*

| | Full cohort | | Not stratified | | | | Stratified | | | |
| | | | $m = 1$ | | $m = 3$ | | $m = 1$ | | $m = 3$ | |
	Ave	SE	Ave	SE	Ave	SE	Ave	SE	Ave	SE
$\widehat{\beta}$	0.483	0.049	0.510	0.133	0.455	0.089	0.508	0.110	0.492	0.080
$c = \widehat{F}_Y^{-1}(0.25)$										
$\widehat{\text{TPF}}_t(c)$	0.900	0.011	0.913	0.027	0.891	0.022	0.906	0.025	0.904	0.017
$\widehat{\text{FPF}}_t(c)$	0.749	0.008	0.748	0.049	0.739	0.030	0.760	0.049	0.787	0.027
$\widehat{\text{PPV}}_t(c)$	0.044	0.004	0.053	0.008	0.053	0.007	0.053	0.006	0.051	0.005
$\widehat{\text{NPV}}_t(c)$	0.985	0.002	0.984	0.005	0.981	0.004	0.982	0.004	0.979	0.004
$c = \widehat{F}_Y^{-1}(0.75)$										
$\widehat{\text{TPF}}_t(c)$	0.471	0.025	0.473	0.057	0.448	0.044	0.482	0.043	0.462	0.033
$\widehat{\text{FPF}}_t(c)$	0.242	0.008	0.225	0.045	0.228	0.028	0.243	0.041	0.272	0.029
$\widehat{\text{PPV}}_t(c)$	0.070	0.007	0.088	0.018	0.083	0.014	0.085	0.013	0.074	0.009
$\widehat{\text{NPV}}_t(c)$	0.974	0.003	0.970	0.005	0.968	0.005	0.969	0.004	0.966	0.004

relatively larger using NCC samples, are still sufficiently precise for making decisions regarding the discriminatory capacity of CRP. For example, with a 24% FPF, we can detect about 47% of the subjects who experience a CVD event or death by 5 years, with a CI of (42%, 52%) estimated from the full cohort. The same detection rate with a CI of (36%, 58%) is observed from an NCC study selecting one matched control per case. Both suggest a significantly better performance than that of a noninformative marker, which is expected to have a TPF of only 24% at this threshold. Such information would be helpful for investigators to plan for more cost-effective future biomarker studies. Based on the results listed in Table 4, we conclude that the predictive accuracy of CRP is quite moderate and that using CRP for preventive decision making should be done with caution. Further assessment of the incremental value of CRP over conventional risk factors is warranted.

6. REMARKS

The NCC sampling has been recognized as a useful design option within cohort study in the field of biomarker research (Rundle *and others*, 2005). Biomarker studies are often influenced by factors such as analytic batch, long-term storage, and freeze-thaw cycles. By matching individual controls to cases' failure times and other potential confounding factors, the accuracy of biomarkers can be evaluated more efficiently and rigorously with an NCC design. However, matching generates complex data that can be more difficult to analyze. We proposed estimators for 2 classes of accuracy measures under an NCC design based on the IPW approach. By reusing samples in each risk set, the approach yields more efficient estimators for those accuracy summaries compared to estimators derived from a partial likelihood, without requiring more information from the study. Furthermore, compared with a nonparametric MLE-based approach (Scheike and Martinussen, 2004), this approach is very simple to implement and is robust to marker-dependent censoring. Extension of the proposed methods to models with multiple covariates and an estimation of covariate-specific accuracy is straightforward. Our theoretical development on the IPW estimators is useful for making inference on accuracy summary estimators in practice, given that justification for bootstrap-based variance estimators has not been developed under cohort sampling settings.

In our estimation procedure, we focused on weights with "true" selection probability according to the study design. The weights can be less optimal. Less expensive covariate information is often available for all subjects in a cohort. Incorporating such information may well yield estimators with improved efficiency. In particular, it is possible to derive semiparametric efficient estimators in this setting following the work of Robins *and others* (1994). Alternatively, parallel to recent work for case–cohort design (Breslow and Wellner, 2007; Breslow *and others*, 2009), one may consider nonparametric procedures to obtain augmented weights using auxiliary information in order to gain efficiency. Our proposal here lays important groundwork for further investigation.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://www.biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

FUNDING

REFERENCES

AKRITAS, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics* **22**, 1299–1327.

ANDERSEN, P. K. AND GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100–1120.

BORGAN, Ø., GOLDSTEIN, L. AND LANGHOLZ, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Annals of Statistics* **23**, 1749–1778.

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. AND KULICH, M. (2009). Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology* **169**, 1398.

BRESLOW, N. E. AND WELLNER, J. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics* **34**, 86–102.

CHEN, K. (2001). Generalized case-cohort sampling. *Journal of the Royal Statistical Society, Series B, Methodological* **63**, 791–809.

GOLDSTEIN, L. AND LANGHOLZ, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Annals of Statistics* **20**, 1903–1928.

HEAGERTY, P. J., LUMLEY, T. AND PEPE, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

HEAGERTY, P. J. AND ZHENG, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105.

Johnson, S. R., Anderson, G. L., Barad, D. H. and Stefanick, M. L. (1999). The women's health initiative: rationale, design, and progress report. *Journal of the British Menopause Society* **5**, 155–159.

Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J. and Castelli, W. P. (1979). An investigation of coronary heart disease in families. The Framingham Offspring Study. *American Journal of Epidemiology* **110**, 281–290.

Langholz, B. and Borgan, Y. (1995). Counter-matching: a stratified nested case-control sampling method. *Biometrika* **82**, 69–79.

Langholz, B. and Borgan, Y. (1997). Estimation of absolute risk from nested case-control data. *Biometrics* **53**, 767–774.

Liang, H. and Baek, J. (2006). Weighted sums of negatively associated random variables. *Australian and New Zealand Journal of Statistics* **48**, 21–31.

Liang, H., Zhang, D. and Baek, J. (2004). Convergence of weighted sums for dependent random variables. *Journal of the Korean Mathematical Society* **41**, 883–894.

Lin, D. Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* **87**, 37.

Lloyd-Jones, D. M., Nam, B. H., D'Agostino, R. B., Levy, D., Murabito, J. M., Wang, T. J., Wilson, P. W. F. and O'Donnell, C. J. (2004). Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults a prospective study of parents and offspring. *Journal of the American Medical Association* **291**, 2204–2211.

Moskowitz, C. and Pepe, M. (2004). Quantifying and comparing the accuracy of binary biomarkers when predicting a failure time outcome. *Statistics in Medicine* **23**, 1555–1570.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W. and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic or prognostic marker. *American Journal of Epidemiology* **159**, 882–890.

Prentice, R. L. and Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65**, 153–158.

Ridker, P. M., Hennekens, C. H., Buring, J. E. and Rifai, N. (2000). C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women. *New England Journal of Medicine* **342**, 836.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Rundle, A. G., Vineis, P. and Ahsan, H. (2005). Design options for molecular epidemiology research within cohort studies. *Cancer Epidemiology Biomarkers & Prevention* **14**, 1899.

Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* **84**, 379–394.

Scheike, T. H. and Martinussen, T. (2004). Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scandinavian Journal of Statistics* **31**, 283–293.

Thomas, D. C. (1977). Addendum to "Methods of cohort analysis: Appraisal by application to asbestos mining". *Journal of the Royal Statistical Society, Series A, General* **140**, 483–485.

Zheng, Y., Cai, T., Pepe, M. S. and Levy, W. C. (2008). Time-dependent predictive values of prognostic biomarkers with failure time outcome. *Journal of the American Statistical Association* **103**, 362–368.

Zheng, Y., Cai, T., Stanford, J. L. and Feng, Z. (2010). Semiparametric models of time-dependent predictive values of prognostic biomarkers. *Biometrics* **66**, 50–60.