Published in final edited form as:

Biometrics. 2012 December ; 68(4): 1219–1227. doi:10.1111/j.1541-0420.2012.01787.x.

Evaluating the Predictive Value of Biomarkers with Stratified Case-Cohort Design

Dandan Liu*, Tianxi Cai†, and Yingye Zheng*

*Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109

†Department of biostatistics, Harvard School of Public Health, Boston, MA 02115

Summary

Identification of novel biomarkers for risk assessment is important for both effective disease prevention and optimal treatment recommendation. Discovery relies on the precious yet limited resource of stored biological samples from large prospective cohort studies. Case-cohort sampling design provides a cost-effective tool in the context of biomarker evaluation, especially when the clinical condition of interest is rare. Existing statistical methods focus on making efficient inference on relative hazard parameters from the Cox regression model. Drawing on recent theoretical development on the weighted likelihood for semiparametric models under two-phase studies (Breslow and Wellner, 2007), we propose statistical methods to evaluate accuracy and predictiveness of a risk prediction biomarker, with censored time-to-event outcome under stratified case-cohort sampling. We consider nonparametric methods and a semiparametric method. We derive large sample properties of proposed estimators and evaluate their finite sample performance using numerical studies. We illustrate new procedures using data from Framingham Offspring study to evaluate the accuracy of a recently developed risk score incorporating biomarker information for predicting cardiovascular disease.

Keywords

Case Cohort Sampling; Negative predictive value; Positive predictive value; Receiver Operating Characteristics Curve (ROC curve); Integrated Discrimination Improvement (IDI); Risk prediction; Survival analysis; Two-phase study

1. Introduction

Novel biomarkers and predictive models to improve risk assessment are important for both effective disease prevention and optimal treatment recommendation. Validating the clinical utility of such predictive markers however requires information from large cohort studies with patients followed prospectively for important clinical outcome. An array of large prospective cohorts has been assembled over the past decade, including, for example, the Atherosclerosis Risk in Communities (ARIC) study (Folsom et al., 2002), the Nurses Health Study (NHS) (Colditz et al., 1997) and the Womens Health Initiative (WHI) (Anderson et al., 2003). With detailed information on epidemiological risk factors and biological samples being collected from cohort members and stored for future molecular marker evaluation, such studies provide valuable research opportunities for risk assessment research.

*yzheng@fhcrc.org.

Supplementary Material

Web Appendices referenced in Section 3 are available with this paper at the *Biometrics* web set on Wiley Online Library.

Efficient sampling and analytical methods are important in such settings. Stored biological samples are a precious yet limited resource that must be used with great care. Assaying markers for the entire cohort can be quite expensive, especially if the clinical condition of interest is rare. It is well known that case-cohort (CCH) sampling (Prentice, 1986), which selects all “cases” and a random subset (known as “subcohort”) of the entire baseline cohort, offers a cost-effective option over full cohort analyses. Originally proposed for assessing exposure-disease association in an assembled cohort, the design has greater value when adopted for molecular marker studies within many of the prospective cohort studies. For example, Folsom et al. (2006) conducted a study aim to assess the incremental coronary risk prediction using C-reactive Protein (CRP) and other risk markers with the ARIC Study cohort. A case-cohort design was adopted for risk marker measurement, involving all available incident CHD cases at the time, and a stratified random sample of the ARIC Study cohort from strata defined by age, sex, and ultrasonographically assessed carotid intima media thickness. A similar sampling approach was used in studying CRP in relation to various cancer outcomes within the WHI cohort (Wang et al., 2011).

The sampling scheme used in the ARIC study is dependent on patients’ disease outcome. Therefore it introduces a complex data structure that makes analysis of case-cohort data challenging. Prior literature has focused on making inference on relative hazard parameters under the Cox regression model, for example using pseudo-likelihood approach for estimating regression parameters under the original case-cohort design (Prentice, 1986) and extending more flexible exposure stratified case-cohort (sCCH) sampling (Borgan et al., 2000). One popular approach in recent years has considered the class of inverse probability weighting (IPW) estimators, where individual contributions are weighted inversely proportional to their sampling fractions (Lin, 2000; Breslow and Wellner, 2007; Gray, 2009). In particular, Breslow and Wellner (2007) derived key theoretical results for estimates of hazard ratios and cumulative event probabilities in the population under Cox regression with sCCH sampling. However, evaluation of the clinical utility of biomarkers, such as the incremental prediction capacity of CRP for CHD as in the ARIC study, adds another level of complexity, and appropriate statistical methods for biomarker assessment and guidance in efficiently conducting the design are still unavailable to the biomarker field.

In biomarker studies it is now well recognized that it is important to evaluate the predictive value of a novel marker/model in multiple ways (Lloyd-Jones, 2010). Specifically, for a putative marker or a predictive score Y measured at baseline and used to predict the occurrence time of a future event T , two classes of time-dependent accuracy parameters, retrospective and prospective accuracies, are commonly calculated as summaries of discrimination. The retrospective accuracies, such as the true and false positive fraction (TPF and FPF), evaluated at a future predicting time t , and a threshold c , are defined as

$$\text{TPR}_t(c) = P(Y > c | T \leq t), \text{ and } \text{FPR}_t(c) = P(Y > c | T > t); \quad (1.1)$$

The time-dependent receiver operating characteristic (ROC_t) curve is a plot of $\text{TPR}_t(c)$ versus $\text{FPR}_t(c)$, across all possible values of c , and the time-dependent area under the ROC curve (AUC_t) summarizes the discriminatory capacity of the marker. The prospective accuracy summaries, positive and negative predictive values (PPV and NPV), are time-dependent functions of the form

$$\text{PPV}_t(c) = P(T \leq t | Y > c), \text{ and } \text{NPV}_t(c) = P(T > t | Y \leq c). \quad (1.2)$$

For a risk marker Y , medical decisions are typically made based on the absolute risk of a poor outcome by t , $\text{Risk}_t(Y) = P(T \leq t | Y)$. Net Reclassification Improvement (NRI) and Integrated Discrimination Improvement (IDI) were first introduced by Pencina et al. (2008)

for assessing improvement in risk reclassification in the context of comparing two risk models constructed with and without novel markers. In the setting of a single marker or a model score, we note here that the key quantity in IDI is the mean risk among the cases minus the mean risk among the controls. We propose to use such a difference in mean risks (DMR) as a measure associated with reclassification:

$$\text{DMR}_t = E\{\text{Risk}_t(Y)|T \leq t\} - E\{\text{Risk}_t(Y)|T > t\}. \quad (1.3)$$

DMR_t measures how well a marker/risk prediction model stratifies patients in terms of risk. It is related to the accuracy measures such that $E\{\text{Risk}_t(Y)|T \leq t\} = \int \text{TPR}_t(c_p) dp$ and $E\{\text{Risk}_t(Y)|T > t\} = \int \text{FPR}_t(c_p) dp$, if $\text{Risk}_t(Y)$ is the true risk and c_p is the threshold value corresponding to $\text{Risk}_t(c_p) = p$.

Statistical models for estimating these quantities using time-to-event data from full cohorts have been proposed (Heagerty, Lumley, and Pepe, 2000; Heagerty and Zheng, 2005; Moskowitz and Pepe, 2004; Zheng, Cai, Pepe, and Levy, 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 the increasing number of new biomarkers for risk prediction, there is an urgent need to expand the research to accommodate case-cohort sampling designs.

In this manuscript we consider statistical procedures for validating the clinical utility of biomarkers based on general cohort sampling study designs. Specifically, we propose IPW-based estimators for estimating key clinical parameters used for quantifying the predictive capacity of novel markers. We describe both nonparametric and semiparametric estimating procedures in such settings. Drawing on the theoretical development in Breslow and Wellner (2007), we further develop asymptotic theories on these proposed estimators in order to facilitate inference making in practice. The paper is organized as follows: we present estimation procedures under a sCCH design in Section 2; asymptotic theory is given in Section 3; Section 4 contains numerical studies. We illustrate sCCH designs and our methods with a study aiming to assess the incremental value of CRP for cardiovascular disease risk prediction in Section 5, and we conclude with some remarks in Section 6.

2. Estimation

2.1 Notation and General Results

Suppose the full cohort consists of N subjects followed prospectively in a clinical study with time-to-event outcomes. For subject i , let T_i , C_i and $X_i = \min(T_i, C_i)$ denote the failure, censoring and observed follow-up times respectively. Let $\Delta_i = I(T_i < C_i)$ be the event indicator and Y_i denote the biomarker variable on which the risk of failure prediction will be based. For sCCH studies, Y_i is not observed for all N subjects. Instead, some surrogate variables $\tilde{Y}_i$ are fully observed, which together with other observed information $\mathbf{Z}_i$ determine the sampling strata. We denote by $\mathcal{H} = \xi\{\mathbf{H}_i = (X_i, \Delta_i, Y_i, \mathbf{Z}_i), i = 1, \dots, n\}$, the sigma field of the information that potentially available for all cohort but observed only for sub cohort members (referred to as the complete full cohort data), and $\tilde{\mathcal{H}} = \xi\{(X_i, \Delta_i, \tilde{Y}_i, \mathbf{Z}_i), i = 1, \dots, n\}$ the observed full cohort data. With V_i being a binary indicator denoting whether subject i is selected in the subcohort, we impose the missing at random (MAR) assumption: $\Pr(V_i = 1|\mathcal{H}) = \Pr(V_i = 1|\tilde{\mathcal{H}}) = \pi_i$, which specifies that inclusion of the subcohort dose not depend on Y given the observed information.

Throughout this paper, we consider a general form of cohort sampling design, assuming subjects are sampled with finite population-stratified sampling as in the case of the ARIC

CRP study described in the Introduction. Suppose given $\mathcal{H}$, L strata are defined, based on surrogate covariates and the case/non-case status determined by the event of interest at the time the sample is drawn. For the l th stratum, n_l subjects are randomly sampled without replacement to have biomarker Y assayed. Such a finite population sampling method is more often implemented in practice than a Bernoulli sampling. Let R_l index the subset of full cohort members who belong to the stratum l , $N_l = |R_l|$, and $v_l = \Pr(i \in R_l) = N_l/N$, $l = 1, \dots, L$. Under the previously specified MAR assumption, the true sampling probability for the i th subject from the l th stratum, π_i , is simply n_l/N_l under a sCCH design.

Let $F(c)$, $S(t)$ and $S(t, c)$ denote the marginal cumulative distribution function of Y , the marginal survival function of T , and the bivariate survival function of T and Y respectively. Simple derivation using Bayes' theorem shows that the aforementioned accuracy measures for Y can be expressed as functions of the above three quantities, namely

$$\text{TRP}_t(c) = \frac{\{1 - F(c)\} - S(t, c)}{1 - S(t)}, \text{FPR}_t(c) = \frac{S(t, c)}{S(t)}, \text{PPV}_t(c) = \frac{\{1 - F(c)\} - S(t, c)}{1 - F(c)}, \text{NPV}_t(c) = \frac{S(t) - S(t, c)}{F(c)}. \quad (2.1)$$

Many other accuracy measures are functions of the above measures. For example,

$\text{ROC}_t(c) = \text{TRP}_t\{\text{FPR}_t^{-1}(c)\}$, $\text{AUC}_t = \int_0^1 \text{ROC}_t(c) dc$, $\text{DMR}_t = \int \{\text{TPR}_t(c_p) - \text{FPR}_t(c_p)\} dp$, where c_p is the threshold value corresponding to $\text{Risk}_t(c_p) = p$. Therefore, consistent estimators for these accuracy measures can be obtained by replacing $F(c)$, $S(t)$ and $S(t, c)$ with their corresponding consistent estimators in the above expression. Without loss of generality, we assume that Y is bounded with finite support $\mathcal{B}_Y = [c_b, c_u]$. Since $S(t) = \Pr(T > t) = \Pr(T > t, Y > c) = S(t, c)$ and $F(c) = \Pr(T > c) = S(0, c)$, one need only direct attention to constructing consistent estimators for $S(t, c)$.

Under a sCCH design, since Y is selectively measured depending on the outcome and other covariates, it is crucial to adjust for the sampling scheme in order to make valid inference about the aforementioned accuracy summaries. We will achieve this by casting the problem into the general framework of failure-time regression with missing covariates and develop IPW estimators for $S(t, c)$. Specifically, we consider weighing the contributions from the i th

observation with weight $w_i = V_i/\pi_i$. In the simplest case $\pi_i = \sum_{l=1}^L I(i \in R_l) n_l / N_l$, which is the true sampling fraction dictated by the sCCH design. Other weights can be considered. For example, Qi et al. (2005) proposed to use $\hat{\pi}$ estimated as a smoothed function of $\mathcal{H}$. Such estimated weight has been shown to improve efficiency. Below we describe both weighted nonparametric and semiparametric procedures for estimating $S(t, c)$. Under both settings w_i are taken to be the design-based weights for simplicity.

2.2 Nonparametric Methods

We first consider nonparametric estimation for $S(t, c)$. Calculating predictive values of a biomarker without imposing any modeling assumption is often desired in biomarker validation study. Nonparametric methods are thus generally applicable to many studies in practice. We develop two sets of estimators varying upon whether marker-dependent or marker-independent censoring is being considered.

Marker-Dependent Censoring—In many biomarker studies, the follow-up efforts may be related to the knowledge of biomarker values measured at the baseline. This may introduce a type of marker-dependent censoring, where censoring time C may depend on marker values Y , while given $Y = y$, the event time T and C are independent. Estimating $S(t, c)$ by a Kaplan-Meier estimator for a subset of subjects with $Y > c$ may introduce bias under such an informative censoring (Murray and Tsiatis, 1996). In a full cohort study, we have

previously considered estimating the bivariate function as $S(t, c) = \int_c^{c_u} S_y(t) dF(y)$, where $S_y(t) = \Pr(T \geq t | Y = y)$ is the conditional survival function given $Y = y$ (Zheng et al., 2008). Here we propose to modify the kernel-smoothed estimator (Beran, 1981; Dabrowska, 1997) for $S_y(t)$ with IPW to account for the stratified case-cohort sampling.

$$\tilde{S}_y(t) = \exp\{-\tilde{\Lambda}_y(t)\} = \exp\left\{-\int_0^t \frac{d\tilde{N}_y(u)}{\tilde{e}_y(u)}\right\}, \quad (2.2)$$

where $\tilde{N}_y(t) = N^{-1} \sum_{i=1}^N w_i K_h(Y_i - y) I(X_i \leq t) \Delta_i$, $\tilde{e}_y(t) = N^{-1} \sum_{i=1}^N w_i K_h(Y_i - y) I(X_i \geq t)$, $K_h(x) = K(xh^{-1})h^{-1}$ and K is a known smooth symmetric kernel density function. The bandwidth parameter h is assumed to be of order $\mathcal{O}(N^{-\nu})$ with $\nu \in [1/5, 1/2)$ to ensure the consistency of $\tilde{S}_y(t)$, however to avoid bias in our final accuracy summaries we require the standard undersmoothing assumption with $\nu \in [1/4, 1/2)$. Since $E(w_i \mathcal{H}) = 1$ and the standard kernel estimators are consistent, the IPW kernel estimator of the conditional survival function is consistent. Using an IPW empirical estimator for $F(c)$,

$$\tilde{F}(c) = \frac{\sum_{i=1}^N w_i I(Y_i \leq c)}{\sum_{i=1}^N w_i}, \quad (2.3)$$

one could obtain a plug-in estimator of $S(t, c)$ as

$$\tilde{S}(t, c) = \int_c^\infty \tilde{S}_y(t) d\tilde{F}(y) = \frac{\sum_{i=1}^N \tilde{S}_{y_i}(t) w_i I(Y_i > c)}{\sum_{i=1}^N w_i}. \quad (2.4)$$

Under marker-dependent censoring, the nonparametric estimators of the accuracy measures $\text{TPR}_t(c)$, $\text{FPR}_t(c)$, $\text{PPV}_t(c)$, $\text{NPV}_t(c)$, AUC_t and DMR_t can be obtained based on (2.3) and (2.4). This set of estimators based on kernel smoothing is termed as SMTH hereafter.

Marker-Independent Censoring—In some studies, censoring is solely due to administrative reasons, thus C can be assumed to be independent of both T and Y . We could construct another set of consistent estimators for the accuracy measures by adding another IPW accounting for missingness due to censoring. We term this set of nonparametric estimators as DIPW (double IPW). Specifically, let $\bar{G}(t)$ denote the Kaplan-Meier estimator of the survival function for the censoring time C , $\bar{G}(t) = \Pr(C \geq t) = \Pr(C \geq t | Y)$. Based on the observed time X_i , the IPW for censoring can be written as $\bar{w}_{C_i}(t) = \{\Delta_i I(X_i \geq t) + I(X_i > t)\} / \bar{G}(X_i \wedge t)$ with $E\{\bar{w}_{C_i}(t) | X_i, \Delta_i, Y_i\} \approx 1$, which is a function of the prediction time t .

The DIPW estimators for $F(c)$ and $S(t, c)$ are then

$$\bar{F}(c) = \frac{\sum_{i=1}^N w_i \bar{w}_{C_i}(t) I(Y_i \leq c)}{\sum_{i=1}^N w_i \bar{w}_{C_i}(t)}, \quad (2.5)$$

$$\bar{S}(t, c) = \frac{\sum_{i=1}^N w_i \bar{w}_{C_i}(t) I(X_i > t, Y_i > c)}{\sum_{i=1}^N w_i \bar{w}_{C_i}(t)}. \quad (2.6)$$

Consequently, the corresponding DIPW estimator for the accuracy measures $\bar{\text{TPR}}_t(c)$, $\bar{\text{FPR}}_t(c)$, $\bar{\text{PPV}}_t(c)$, $\bar{\text{NPV}}_t(c)$, $\bar{\text{AUC}}_t$ and $\bar{\text{DMR}}_t$ can be obtained based on (2.5) and

(2.6). It should be noted that we use the double weights for $\bar{R}(c)$ to ensure the estimated accuracy measures are between 0 and 1.

Both sets of nonparametric estimators are useful in practice. SMTH estimators are more flexible and may be more efficient, whereas DIPW estimators are simple to implement and may produce more stable performance in small sample size.

2.3 Semiparametric Methods

The Cox (1972) proportional hazards model: $\lambda(t|Y=y) = \lambda_0(t) \exp(\beta_0 Y)$ is the most widely used semiparametric model for time-to-event endpoints, where $\lambda_0(t)$ is an unspecified function and β_0 is the regression parameter for Y . With outcome-dependent stratified sampling, the IPW estimators for β_0 was discussed by Borgan et al. (2000, estimator II). Breslow and Wellner (2007) demonstrated the consistency and weak convergence of the IPW estimators for $\lambda_0(t)$ and β_0 . Based on their results, we propose to estimate the conditional survival function as $\hat{S}_y(t) = \exp\{-\hat{\Lambda}_0(t) \exp(\hat{\beta}_y)\}$, where

$$\hat{\Lambda}_0(t) = \sum_{i=1}^N \frac{w_i I(X_i \leq t) \Delta_i}{\sum_{j=1}^N I(X_j \geq X_i) w_i \exp(\hat{\beta}_Y)}$$

is the IPW Breslow estimator for $\Lambda_0(t)$. Together with (2.3), the semiparametric estimator for the bivariate survival function is then

$$\hat{S}(t, c) = \int_c^\infty \hat{S}_y(t) d\tilde{F}(y) = \sum_{i=1}^N \hat{S}_{y_i}(t) w_i I(Y_i > c) \bigg/ \sum_{i=1}^N w_i. \quad (2.7)$$

Subsequently, the semiparametric estimator (SEMI) for the accuracy measures $\widehat{\text{TPR}}_t(c)$, $\widehat{\text{FPR}}_t(c)$, $\widehat{\text{PPV}}_t(c)$, $\widehat{\text{NPV}}_t(c)$, $\widehat{\text{AUC}}_t$ and $\widehat{\text{DMR}}_t$ can be obtained based on (2.3) and (2.7).

3. Statistical Inference

In this section, we outline theoretical results for SMTH, DIPW and SEMI estimators with design-based sampling weights. We note that the main challenge in establishing the asymptotic results comes from the negative correlation between the selection indicators for two individuals from the same stratum I , V_j and $V_{j'}$ with $j \in R_I$ and $j' \in R_I$ due to finite population stratified sampling. The standard weak convergence theory does not apply here because of the correlation. The derivation of asymptotic properties of our proposed

estimators $\mathbb{P}_N^\pi = N^{-1} \sum_{i=1}^N w_i \delta_{\mathbf{H}_i}$ follows directly from the weak convergence of the probability measure to $\mathbb{P}_0$ (Breslow and Wellner, 2007), where $\delta_{\mathbf{H}_i}$ denotes Dirac measure placing unit mass at $\mathbf{H}_i$ and $\mathbb{P}_0$ is the true distribution of $\mathbf{H}$. In Breslow and Wellner (2007), the process $\sqrt{N}(\mathbb{P}_N^\pi - \mathbb{P}_0)$ was shown to converge weakly to a mean zero Gaussian process consisting of two independent components, one accounting for the variation from the full cohort data $\mathcal{H}$ and the other accounting for the variation due to sampling. This convergence along with standard empirical process theory (Pollard, 1990) can be used to obtain the consistency and weak convergence of our proposed estimators. The regularity conditions required to ensure the convergence are listed in Appendix A of Web supplementary material. We next summarize the theoretical results for the proposed estimators, beginning with the following Lemma.

LEMMA 1: Under the regularity conditions (a)-(f) in Appendix A, $\tilde{F}(c)$, $\tilde{\Lambda}_Y(t)$, $\tilde{S}_Y(t)$, $\tilde{S}(t, c)$, $\tilde{S}(t, c)$, $\hat{\beta}$, $\hat{\Lambda}_0(t)$ and $\hat{S}(t, c)$ are consistent estimators.

With the consistency of the above estimators, we show in Appendix B that for any of the accuracy measures, $\text{FPR}_\lambda(c)$, $\text{TPR}_\lambda(c)$, $\text{PPV}_\lambda(c)$ and $\text{NPV}_\lambda(c)$, denoted by a generic term $A_\lambda(c)$, and AUC_t and DMR_t by A_b , the corresponding empirical processes using the SMTH, DIPW and SEMI methods can be written as summations over N subjects as respectively, where

$$\begin{aligned}\tilde{W}_{A_t}(c) &= N^{\frac{1}{2}} \{ \tilde{A}_t(c) - A_t(c) \} = N^{-\frac{1}{2}} \sum_{i=1}^N w_i \tilde{U}_{A_t}(c; \mathcal{H}_i) + o_p(1), \\ \bar{W}_{A_t}(c) &= N^{\frac{1}{2}} \{ \bar{A}_t(c) - A_t(c) \} = N^{-\frac{1}{2}} \sum_{i=1}^N w_i \bar{U}_{A_t}(c; \mathcal{H}_i) + o_p(1), \\ \hat{W}_{A_t}(c) &= N^{\frac{1}{2}} \{ \hat{A}_t(c) - A_t(c) \} = N^{-\frac{1}{2}} \sum_{i=1}^N w_i \hat{U}_{A_t}(c; \mathcal{H}_i) + o_p(1),\end{aligned}$$

$\tilde{U}_{A_t}(c; \mathcal{H}_i)$, $\bar{U}_{A_t}(c; \mathcal{H}_i)$ and $\hat{U}_{A_t}(c; \mathcal{H}_i)$ are functions of the complete full cohort data. Similarly, we also show in Appendix B the corresponding asymptotic expansion for $\tilde{W}_{A_r}$, $\bar{W}_{A_r}$ and $\hat{W}_{A_r}$.

To construct confidence intervals for our proposed estimators, we summarize the asymptotic normality results in the following theorems.

THEOREM 1: Under the regularity conditions (a)-(e), $N^{\frac{1}{2}} \{ \tilde{A}_t(c) - A_t(c) \}$ converges weakly to a Gaussian process in c with mean zero and variance function

$\Sigma_{\tilde{A}_t}(c) = Q\{\tilde{U}_{A_t}(c)\}$ and $N^{\frac{1}{2}} \{ \tilde{A}_t - A_t \}$ converges weakly to a Gaussian distribution with mean zero and variance function $\Sigma_{\tilde{A}_t} = Q\{\tilde{U}_{A_t}\}$, where $Q(U) = \text{var}(U) + \sum_{l=1}^L v_l(\pi_l^{-1} - 1)\text{var}_l(U)$ with var_l denoting the variance within the l th stratum.

THEOREM 2: Under the regularity conditions (a)-(c), $N^{\frac{1}{2}} \{ \bar{A}_t(c) - A_t(c) \}$ converges weakly to a Gaussian process in c with mean zero and variance function $\Sigma_{\bar{A}_t}(c) = Q\{\bar{U}_{A_t}(c)\}$; $N^{\frac{1}{2}} \{ \bar{A}_t - A_t \}$ converges weakly to a Gaussian distribution with mean zero and variance function $\Sigma_{\bar{A}_t} = Q\{\bar{U}_{A_t}\}$.

THEOREM 3: under the regularity conditions (a)-(d) and (f), $N^{\frac{1}{2}} \{ \hat{A}_t(c) - A_t(c) \}$ converges weakly to a Gaussian process in c with mean zero and variance function

$\Sigma_{\hat{A}_t}(c) = Q\{\hat{U}_{A_t}(c)\}$; $N^{\frac{1}{2}} \{ \hat{A}_t - A_t \}$ converges weakly to a Gaussian distribution with mean zero and variance function $\Sigma_{\hat{A}_t} = Q\{\hat{U}_{A_t}\}$.

To provide some insights and intuition behind the general theory, we outline the derivation of our asymptotic variance forms in Appendix C. The key is to account for the covariance between V_{ji} and $V_{j'1}$ in the calculation. We also show that ignoring the covariance between V_{ji} and $V_{j'1}$ leads to the popular robust variance for the original case-cohort design.

4. Simulation Study for the Accuracy Measures

We conducted simulations to assess the performance of the proposed inference procedure. The continuous biomarker variable Y is generated from truncated normal such that $Y = \text{sign}(\tilde{Y}) \min(|\tilde{Y}|, 5)$ with $\tilde{Y} \sim N(0, 1)$. We then simulate failure time from Cox model

$$\lambda(t|Y) = \lambda_0(t) \exp(\beta Y) \quad (4.1)$$

with $\lambda_0(t) = 0.1$ and $\beta = \log(3)$. We let the censoring time $C = \min(C_0, 2)$ with $C_0 \sim \text{Gamma}(2.5, 2)$, resulting in about 84% of censoring. The stratum are defined with case/noncase status and a binary covariate Z , with $Z = I(Y > 0)$, representing a surrogate measure of Y . The four strata, control with $Y = 0$, control with $Y > 0$, case with $Y = 0$ and case with $Y > 0$, each has a stratum proportion of 46.8%, 37.8%, 2.9% and 12.5%, respectively. We consider a full cohort of size $N = 8000$ and sample 200 subjects without replacement from each stratum. For each simulated dataset, we calculate the estimators and the corresponding standard errors for different summary measures of Y in predicting the risk of having an event by $t_0 = 1$ using both the nonparametric methods (Table 1) and the semiparametric (SEMI) method (Table 2). We then summarize results based on 1000 simulated datasets.

We present results for $\text{FPR}_\lambda(c)$, $\text{TPR}_\lambda(c)$, $\text{NPV}_\lambda(c)$ and $\text{PPV}_\lambda(c)$ at threshold values taken to be the 25th, 50th and 75th percentile of Y , i.e., $c = \mathcal{F}^{-1}(p)$ for $p = 0.25, 0.5, 0.75$. (Table 1(a) and 2(a)), and results for global measures AUC and MDR, as well as FPR, NPV and PPV at a TPR level of 0.9 (Table 1(b) and 2(b)). It appears that all our proposed estimators are unbiased. Our asymptotic standard errors (ASE) track their empirical standard deviations (ESD) closely, and the coverage probabilities (CP) are closed to the nominal value 95%. We therefore concluded that our proposed nonparametric and semiparametric estimators and their variance estimators perform well in finite samples. In contrast, all the robust variance based estimators have ASEs that tend to overestimate ESD, in many cases yielding CPs close to 1. We therefore recommend against using robust variance based estimators for making inference on summary measures with stratified case-cohort sampling for biomarkers.

5. Application

The Framingham Offspring Study was established in 1971 with 5,124 participants who were monitored prospectively on epidemiological and genetic risk factors of CVD. We consider here 1687 female participants who are free of CVD and have CRP measurements and other clinical information from the second examination. The average age of this subset was about 44 years with standard deviation 10. The outcome we consider is the time from exam date to first major CVD event, including CVD-related death, with censoring time at 10 years of follow-up. During the follow-up period, 67 participants experienced at least one CVD event, and the 5-year event rate was about 2%. Different from the ARIC CRP study where biomarkers were only measured for a subset of cohort members, CRP measurements are complete in the Framingham cohort. Therefore the data allow us to illustrate the methods with a real dataset and compare estimators obtained using data from sampled subcohorts to those from the full cohort.

We first calculated a new risk score using an algorithm developed previously in Cook et al. (2006), combining information on age, systolic blood pressure, smoking status, high-density lipoprotein (HDL), total cholesterol, medication for hypertension and CRP concentration. The score was derived using a Cox proportional hazards model. Since the framingham score is correlated with the risk score, it is a sensible surrogate variable to define stratum for sampling the subcohort. We consider three strata, the cases, the controls with framingham scores larger than the median and the controls with framingham scores less than or equal to median. For the case-cohort design, we keep all the cases and sample 1.5 times the cases from each control stratum, resulting a case-control ratio of 1:3 and a sample size of 268, i.e., less than 16% of the full cohort sample size. We assembled 500 case-cohort datasets by repeatedly sampling the controls. For each dataset, we obtained the point and standard error estimates of accuracy summaries for the risk score in predicting the risk of developing CVD

events within 5 years since predictor measurements. As shown in Table 4, the new risk score appeared to have moderate capacity to discriminate individuals who would experience a CVD events in the next 5 years from those who would not, with an estimated AUC of 0.751 (SE = 0.042) from the full cohort data, and AUC_c of 0.750 (SE = 0.044) for case-cohort samples based on the semiparametric method. The score also appears to have good performance for stratifying patients in terms of 5-year risks, with a DMR evaluated at 5 years of 0.255 (SE = 0.032) based on the semiparametric method for the sCCH design. In Figure 1 we present ROC curves and PPV curves calculated with both semiparametric and nonparametric methods. Estimated ROC curves and PPV curves from the case-cohort samples trace closely their counterparts from full cohort data. Most impressively, for this study the loss in efficiency from using the subcohort sampling design appears to be ignorable for both sets of accuracy summary measures and across semiparametric and nonparametric methods, albeit only 16% of the data is used for estimation. This provides useful insights for designing more efficient biomarker validation studies in the future using the information from the cohort.

6. Remarks

Cost-effective sampling strategies are urgently needed in biomarker field. Sampling from a cohort requires deliberation to avoid selection bias. The case-cohort study design, widely used in epidemiological association studies and clinical trials, has been shown to yield unbiased relative risk estimates and appears to be quite efficient compared with a full cohort analysis. Despite the advantages, the design is used infrequently in the biomarker field, and convenient subcohorts that are prone to all sources of bias in a cohort situation are often selected. Investigators are often unaware of the design and its potential advantage, and more often they may be deterred by the lack of formal statistical methods appropriate for biomarker evaluation in this setting. Existing literature in cohort sampling has been mainly focused on the relative risk estimates, a parameter unsuitable for evaluating marker performance.

In this manuscript, we develop statistical tools that directly address the questions that arise in biomarker evaluation with cohort sampling. We proposed a class of IPW based estimators for an array of measures commonly used in biomarker validation study. Furthermore, we studied their asymptotic properties to aid their inferences in practice, drawing on recent theoretical development for two-phase studies (Breslow and Wellner, 2007). This is an important contribution as resampling based methods such as bootstrap variance estimators are not appropriate for case-cohort studies. We are currently working on developing software with intuitive user interfaces and detailed documentation that will be freely available for biomarker validation studies, and research codes are currently available upon request. We hope that by providing appropriate statistical tools for biomarker study and elucidating key design issues, our research effort will help move the field toward standards of practice.

In our estimation procedure we focused on weights with true selection probability according to the study design. The weights can be less optimal. Furthermore, 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 et al. (1994). Alternatively, following the recent work of Breslow and Wellner (2007) and Breslow et al. (2009), one may consider nonparametric procedures to obtain augmented weights using auxiliary information in order to gain efficiency. Further investigation is warranted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The Framingham Heart Study and the Framingham SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. The Framingham SHARe data used for the analyses described in this manuscript were obtained through dbGaP (access number: phs000007.v3.p2). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI. The work is supported by grants U01-CA86368, P01-CA053996, R01-GM085047 and R01-GM079330 awarded by the National Institutes of Health.

References

- Anderson G, Manson J, Wallace R, Lund B, Hall D, Davis S, Shumaker S, Wang C, Stein E, Prentice R. Implementation of the Women's Health Initiative study design. *Annals of Epidemiology*. 2003; 13:S5. [PubMed: 14575938]
- Beran, R. Unpublished manuscript. Berkeley: Univ. of California; 1981. Nonparametric regression with randomly censored survival data.
- Borgan O, Langholz B, Samuelsen S, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. *Lifetime Data Analysis*. 2000; 6:39–58. [PubMed: 10763560]
- Breslow N, Lumley T, Ballantyne C, Chambless L, Kulich M. Improved Horvitz-Thompson Estimation of Model Parameters from Two-phase Stratified Samples: Applications in Epidemiology. *Statistics in Biosciences*. 2009; 1:32–49. [PubMed: 20174455]
- Breslow N, Wellner J. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scand. J. Statist*. 2007; 34:86–102.
- Colditz G, Manson J, Hankinson S. The nurses' health study: 20-year contribution to the understanding of health among women. *Journal of Women's Health*. 1997; 6:49–62.
- Cook N, Buring J, Ridker P. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Annals of Internal Medicine*. 2006; 145:21. [PubMed: 16818925]
- Cox D. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1972; 34:187–220.
- Dabrowska D. Smoothed cox regression. *The Annals of Statistics*. 1997; 25:1510–1540.
- Folsom A, Aleksic N, Catellier D, Juneja H, Wu K. C-reactive protein and incident coronary heart disease in the atherosclerosis risk in communities (aric) study. *American Heart Journal*. 2002; 144:233–238. [PubMed: 12177639]
- Folsom A, Chambless L, Ballantyne C, Coresh J, Heiss G, Wu K, Boerwinkle E, Mosley T Jr. Sorlie P, Diao G, et al. An assessment of incremental coronary risk prediction using c-reactive protein and other novel risk markers: the atherosclerosis risk in communities study. *Archives of internal medicine*. 2006; 166:1368. [PubMed: 16832001]
- Gray R. Weighted analyses for cohort sampling designs. *Lifetime Data Analysis*. 2009; 15:24–40. [PubMed: 18712477]
- Heagerty P, Lumley T, Pepe M. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000; 56:337–344. [PubMed: 10877287]
- Heagerty P, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics*. 2005; 61:92–105. [PubMed: 15737082]
- Lin D. On fitting cox's proportional hazards models to survey data. *Biometrika*. 2000; 87:37–47.
- Lloyd-Jones D. Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation*. 2010; 121:1768. [PubMed: 20404268]
- Moskowitz C, Pepe M. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics*. 2004; 5:113. [PubMed: 14744831]
- Murray S, Tsiatis A. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics*. 1996; 52:137–151. [PubMed: 8934589]

- Pencina M, D'Agostino R Sr, D'Agostino R Jr, Vasan R. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*. 2008; 27:157–172. [PubMed: 17569110]
- Pollard, D. Empirical processes: theory and applications. Institute of Mathematical Statistics and the American Statistical Association; 1990.
- Prentice R. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Qi L, Wang C, Prentice R. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 2005; 100:1250–1263.
- Robins J, Rotnitzky A, Zhao L. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
- Wang T, Rohan T, Gunter M, Xue X, Wactawski-Wende J, Rajpathak S, Cushman M, Strickler H, Kaplan R, Wassertheil-Smoller S, et al. A prospective study of inflammation markers and endometrial cancer risk in postmenopausal hormone nonusers. *Cancer Epidemiology Biomarkers & Prevention*. 2011; 20:971–977.
- Zheng Y, Cai T, Pepe M, Levy W. Time-dependent predictive values of prognostic biomarkers with failure time outcome. *Journal of the American Statistical Association*. 2008; 103:362–368. [PubMed: 19655041]

Figure 1.

Estimates of prognostic accuracy of a new risk score for 5 year CVD risk from the Framingham offspring study: ROC curves from the semiparametric method (a) and the nonparametric kernel based method (b); and PPV curves from the semiparametric method (c) and the nonparametric kernel based method (d). Solid lines: estimates using the full cohort data; shaded area: 95% pointwise confidence interval for full cohort estimates; dashed lines: estimates and 95% pointwise confidence interval using sCCH sampling.

Table 1

Finite sample performance of the proposed nonparametric SMTH and DIPW estimators. Shown below are bias and empirical standard deviation (ESD) of the proposed estimators, and average of the estimated standard error (ASE) and empirical coverage level of the 95% confidence intervals (CP) for proposed (STR) and robust (ROB) interval estimators.

(a) Accuracy estimates at $c_p = \hat{\sigma}^{-1}(\rho)$ for $p = 0.25, 0.5, 0.75$.													
	SMTH						DIPW						
	True	Bias	ESD	ASE	CP(%)	ROB	STR	ASE	ESD	Bias	CP(%)	ROB	
TPR($c_{2.5}$)	0.953	-0.002	0.008	0.008	95.6	0.008	96.5	0.000	0.007	0.007	93.5	0.008	95.3
TPR(c_5)	0.834	-0.005	0.015	0.014	93.8	0.017	97.3	0.000	0.013	0.013	94.1	0.016	98.2
TPR($c_{7.5}$)	0.602	-0.008	0.030	0.032	95.5	0.033	95.9	0.001	0.031	0.032	95.2	0.032	95.7
FPR($c_{2.5}$)	0.715	0.003	0.019	0.020	95.2	0.023	97.9	0.002	0.029	0.029	94.7	0.031	96.7
FPR(c_5)	0.443	0.002	0.007	0.007	94.6	0.024	100	0.000	0.023	0.023	95.2	0.032	99.2
FPR($c_{7.5}$)	0.190	0.003	0.015	0.016	94.5	0.018	98.0	0.001	0.022	0.022	94.8	0.024	96.9
PPV($c_{2.5}$)	0.185	-0.003	0.010	0.010	93.0	0.014	99.3	0.000	0.012	0.012	95.1	0.018	99.5
PPV(c_5)	0.243	-0.005	0.013	0.013	93.1	0.020	99.3	0.001	0.016	0.017	95.1	0.025	99.7
PPV($c_{7.5}$)	0.350	-0.009	0.027	0.027	92.8	0.032	96.6	0.002	0.034	0.034	94.5	0.040	97.2
NPV($c_{2.5}$)	0.973	-0.001	0.004	0.005	96.4	0.005	97.5	0.000	0.005	0.005	95.4	0.006	97.0
NPV(c_5)	0.952	-0.001	0.004	0.004	94.2	0.005	98.3	0.000	0.005	0.005	95.5	0.007	98.7
NPV($c_{7.5}$)	0.923	-0.001	0.006	0.007	96.9	0.008	98.8	0.000	0.007	0.008	96.1	0.010	99.4
(b) AUC, DMR and accuracy estimates at $\text{TPR}_q = 0.9$.													
	SMTH						DIPW						
	True	Bias	ESD	ASE	CP(%)	ROB	STR	ASE	ESD	Bias	CP(%)	ROB	
AUC _{ℓ}	0.786	-0.006	0.014	0.014	0.935	0.016	97.3	0.000	0.016	0.016	94.8	0.018	97.2
DMR _{ℓ}	0.292	-0.007	0.014	0.014	0.924	0.016	96.5	-0.001	0.018	0.020	97.4	0.024	98.9
FPR _{ℓ} [*]	0.568	0.009	0.029	0.033	0.974	0.039	99.3	-0.002	0.038	0.039	94.8	0.044	98.0
NPV _{ℓ} [*]	0.962	-0.001	0.003	0.003	0.967	0.004	99.4	0.000	0.004	0.004	94.7	0.005	99.1

(b) AUC, DMR and accuracy estimates at $TPR_t = 0.9$.

DIPW											
SMTH						ROB					
STR			ASE			CP(%)			ROB		
True	Bias	ESD	ASE	CP(%)	ASE	CP(%)	ASE	CP(%)	ASE	CP(%)	ROB
PPV _t *	0.212	-0.005	0.014	0.014	0.941	0.020	0.020	99.2	0.001	0.016	0.017
											96.0
											0.024
											99.6

* evaluated at $TPR_t = 0.9$.

Table 2

Finite sample performance of the proposed semiparametric estimators. Shown below are bias and empirical standard deviation (ESD) of the proposed estimators, and average of the estimated standard error (ASE) and empirical coverage level of the 95% confidence intervals (CP) for proposed (STR) and robust (ROB) interval estimators.

(a) Accuracy estimates at $c_p = \hat{\sigma}^{-1}(p)$ for $p = 0.25, 0.5, 0.75$.									
	True	Bias	ESD	STR		ASE	CP(%)	ROB	
				ASE	CP(%)			ASE	CP(%)
TPR _(c₂₅)	0.953	0.000	0.007	0.007	93.0	0.007	94.1		
TPR _(c₅₀)	0.834	0.000	0.015	0.014	93.5	0.017	96.3		
TPR _(c₇₅)	0.602	0.001	0.026	0.025	93.4	0.026	94.1		
FPR _(c₂₅)	0.715	0.002	0.019	0.019	95.1	0.022	98.1		
FPR _(c₅₀)	0.443	0.000	0.007	0.007	94.5	0.023	100		
FPR _(c₇₅)	0.190	0.001	0.014	0.014	95.1	0.017	98.1		
PPV _(c₂₅)	0.185	0.000	0.009	0.009	94.8	0.013	99.6		
PPV _(c₅₀)	0.243	0.000	0.012	0.012	94.8	0.018	99.5		
PPV _(c₇₅)	0.350	0.001	0.012	0.012	94.7	0.017	99.3		
NPV _(c₂₅)	0.973	0.000	0.004	0.004	94.2	0.004	96.0		
NPV _(c₅₀)	0.952	0.000	0.005	0.005	93.8	0.005	97.3		
NPV _(c₇₅)	0.923	0.000	0.003	0.003	94.2	0.004	98.8		
(b) AUC, DMR and accuracy estimates at $\text{TPR}_t = 0.9$.									
	True	Bias	ESD	STR		ASE	CP(%)	ROB	
				ASE	CP(%)			ASE	CP(%)
AUC _t	0.786	0.001	0.014	0.013	93.6	0.015	95.9		
DMR _t	0.292	0.000	0.014	0.013	93.9	0.015	96.6		
FPR _(*)	0.568	-0.002	0.027	0.027	94.4	0.029	95.5		
NPV _(*)	0.962	0.000	0.003	0.003	94.2	0.004	98.8		
PPV _(*)	0.212	0.001	0.012	0.012	94.7	0.017	99.3		

* evaluated at $\text{TPR}_f = 0.9$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

Estimated (Est) accuracy measures for predicting 5-year CVD events and standard errors (SE) from the Framingham Study using data from full cohort (cohort) and a stratified case cohort (sCCH) sampling method.

(a) Accuracy estimates at $c_p = F^{-1}(p)$ for $p = 0.25, 0.5, 0.75$.										
	SMTH			SEMI						
	Cohort		sCCH	Cohort		sCCH				
	Est	SE	Est	SE	Est	SE		Est	SE	
TPR(c_{25})	0.916	0.046	0.919	0.047	0.944	0.014	0.945	0.017		
TPR(c_{50})	0.830	0.062	0.830	0.063	0.831	0.031	0.833	0.037		
TPR(c_{75})	0.690	0.077	0.680	0.078	0.609	0.047	0.611	0.054		
FPR(c_{25})	0.746	0.011	0.747	0.026	0.746	0.011	0.746	0.026		
FPR(c_{50})	0.493	0.012	0.492	0.024	0.492	0.012	0.492	0.024		
FPR(c_{75})	0.241	0.011	0.240	0.024	0.242	0.010	0.241	0.024		
PPV(c_{25})	0.026	0.004	0.026	0.005	0.027	0.004	0.027	0.005		
PPV(c_{50})	0.035	0.006	0.035	0.006	0.036	0.006	0.036	0.006		
PPV(c_{75})	0.058	0.011	0.057	0.013	0.052	0.009	0.053	0.010		
NPV(c_{25})	0.993	0.004	0.993	0.004	0.995	0.001	0.995	0.002		
NPV(c_{50})	0.993	0.003	0.993	0.003	0.993	0.002	0.993	0.002		
NPV(c_{75})	0.991	0.003	0.991	0.003	0.989	0.002	0.989	0.002		
(b) AUC, DMR and accuracy estimates at $TPR_t = 0.9$.										
	SMTH			SEMI						
	Cohort		sCCH	Cohort		sCCH				
	Est	SE	Est	SE	Est	SE		Est	SE	
AUC $_t$	0.757	0.042	0.751	0.044	0.751	0.027	0.750	0.031		
DMR $_t$	0.257	0.042	0.257	0.045	0.251	0.027	0.255	0.032		
FPR(c^*)	0.972	0.027	0.970	0.029	0.985	0.004	0.985	0.005		
PPV(c^*)	0.023	0.004	0.022	0.004	0.023	0.004	0.023	0.004		

(b) AUC, DMR and accuracy estimates at $TPR_f = 0.9$.									
SMTH					SEMI				
Cohort		sCCH			Cohort		sCCH		
Est	SE	Est	SE	SE	Est	SE	Est	SE	SE
NPV _f [*]		0.994	0.006	0.994	0.006	0.997	0.001	0.997	0.001

^{*} evaluated at $TPR_f = 0.9$.