Analysis of Cohort Studies with Multivariate, Partially Observed, Disease Classification Data

By Nilanjan Chatterjee
Division of Cancer Epidemiology and Genetics,
National Cancer Institute, NIH, DHHS. Rockville, MD 20852, USA.
chattern@mail.nih.gov

Samiran Sinha
Texas A&M University, College Station, TX 77843, USA.
sinha@stat.tamu.edu

W. Ryan Diver and Heather Spencer Feigelson
Department of Epidemiology and Surveillance Research,
American Cancer Society, Atlanta, GA 30303, USA.

Summary

Complex diseases like cancer can often be classified into subtypes using various pathological and molecular traits of the disease. In this article, we develop methods for analysis of disease incidence in cohort studies incorporating data on multiple disease traits using a two-stage semiparametric Cox proportional hazards regression model that allows one to examine the heterogeneity in the effect of the covariates by the levels of the different disease traits. For inference in the presence of missing disease traits, we propose a generalization of an estimating-equation approach for handling missing cause of failure in competing-risk data. We prove asymptotic unbiasedness of the estimating-equation method under a general missing-at-random assumption and propose a novel influence-function based sandwich variance estimator. The methods are illustrated using simulation studies and a real data application involving the Cancer Prevention Study (CPS-II) nutrition cohort.

Some key words: Competing-risk; Etiologic heterogeneity; Influence function; Missing cause of failure; Partial likelihood; Proportional hazard regression; Two-stage model.

1. Introduction

Epidemiological researchers commonly use prospective cohort studies to investigate risk factors associated with the *incidence* of chronic diseases such as heart disease, diabetes and cancer. The proportional hazards model (Cox, 1972) is widely used to analyze data from cohort studies for the purpose of making inferences about covariate relative-risk/hazard-ratio parameters. In the standard Cox model, the disease of interest is treated as a single event and the time to incidence of the event is treated as the outcome of interest. In modern epidemiological studies, however, the disease of interest can often be classified into finer subtypes based on various pathologic and molecular traits of the disease. Although there has recently been tremendous progress in methods for using such disease classification data for the study of survival and prognosis after disease incidence, much less attention has been given to methods for incorporating such disease trait data into an etiologic investigation of a disease.

In this article, we develop methods for incorporating disease trait information into the analysis of cohort data with the scientific aim of studying "etiologic heterogeneity", that is, whether the effects of the underlying risk factors vary for the different subtypes of a disease. The basic idea involves using a competing-risk framework to model the hazards of different disease subtypes separately. There are, however, two major analytic complexities. First, a combination of multiple disease traits, some of which can be ordinal or even continuous, can potentially define a very large number of disease subtypes. Using each disease subtype as a separate entity, without imposing any further structure, will require a large number of parameters in the underlying Cox models, potentially causing problems of interpretation, inefficiency, and power loss in related testing procedures due to the imprecision of parameter estimates and the large number of degrees of freedom. Second, the disease classification data in epidemiologic studies can often be incomplete for a large number of subjects due to missing data for the underlying disease traits. A complete-case analysis, using only those diseased subjects who have complete trait

information, can result in both bias and inefficiency.

We propose dealing with these problems by the use of a two-stage regression model coupled with an estimating-equation inferential method. There are several novel aspects to our proposal. First, motivated by our earlier work on polytomous logistic regression models (Chatterjee, 2004), we propose to reduce the number of parameters in the competing-risk proportional hazard model by imposing a natural structure on the relative-risk parameters through the underlying disease traits. The parameters of the reduced model themselves are of scientific interest and are useful for testing etiologic heterogeneity in terms of the underlying disease traits. Second, for the purpose of inference, we propose a general extension of an estimating equation method of Goetghebeur & Ryan (1995) that was originally designed to deal with missing information on causes of failure in an unstructured competing-risk problem. The proposed extension allows one to incorporate the underlying structure of the competing events through a "second-stage" trait design-matrix with the missing data being dealt with by taking suitable expectations of the design-matrix, given the observed covariate and trait data. Third, we prove unbiasedness of our estimating-equation method under a more general missing-at-random assumption than that has been considered before. Finally, we use empirical process theory to develop the asymptotic theory of our estimating equation method and an associated influence-function-based robust sandwich variance estimator under the general missing-at-random assumption. The finite sample properties of the proposed estimator are studied via a simulation studies involving small and large numbers of disease subtypes. Moreover, we apply the proposed method to a data set on breast cancer incidence and various histopathologic traits of breast tumors from the well-known Cancer Prevention Study (CPS-II) of the American Cancer Society.

2. Model and Assumption

Suppose that in a cohort study of size n, each subject is followed until the first occurrence of the disease of interest or the censoring, whichever comes first. Following standard convention, let T denote the underlying time-to-event for the disease and C denote time to censoring. For standard cohort analysis, the outcome is represented by (Δ, V), where $\Delta = I(T < C)$ denotes the indicator of whether or not the disease occurred before censoring and $V = \min(C, T)$ denotes the time-to-censoring or time-to-disease, whichever occurs first. Let us assume that, if disease occurs ($\Delta = 1$), the study collects data on K disease traits, $Y = (Y_1, \ldots, Y_K)$, which could be, for example, various tumor characteristics. If the k-th trait defines M_k categories for the disease, then potentially one can define potentially a total of $M = M_1 \times M_2 \times \ldots M_K$ subtypes, based on all possible combinations of the various characteristics. Breast cancer patients, for example, are often classified based on estrogen- and progesterone-receptor status into four categories: ER+PR+, ER+PR-, ER-PR+ and ER-PR-, where +/- indicates the presence/absence of the corresponding receptor in the breast tumor. Given that follow-up ends at the occurrence of any type of breast cancer, the M subtypes of the disease can be treated as competing events. If X denotes a variate vector of covariates of interest, assumed without loss of generality to be time independent, one can use the proportional hazards model to specify the cause-specific hazard for each subtype of the disease as

$$h_y(t|X) \;=\; \lim_{h\to 0} h^{-1}\mathrm{pr}(t \le T < t+h, Y = y | T \ge t, X) = \lambda_y(t)\exp(X\beta_y),$$

where $\lambda_y(\cdot)$ is the baseline hazard function and β_y is the log-hazard-ratio parameter associated with the disease subtype y. A complication is that modern molecular epidemiologic studies collect data on an array of different traits, which can be represented by a mixture of categorical, ordinal and continuous variable(s). In the above approach, even with few covariates and disease traits, the total number of regression coefficients easily can become very large. In the following section, we consider reducing the number

of parameters by using a second-stage model, following an idea introduced by Chatterjee (2004) in the context of a polytomous logistic regression model.

2.1. *A Second-stage Regression Model for the Subtype-specific Regression Parameter*

First, we focus on modeling the regression coefficients associated with a single covariate. We note that the indexing of the different disease subtypes by the K underlying disease traits immediately suggests the following log-linear representation for the hazard-ratio parameters:

$$\beta_{(y_1,\cdots,y_K)} = \theta^{(0)} + \sum_{k=1}^{K} \theta^{(1)}_{k(y_k)} + \sum_{k=1}^{K}\sum_{k'\geq k}^{K} \theta^{(2)}_{kk'(y_k,y_{k'})} + \cdots + \theta^{(K)}_{12\cdots K(y_1,\cdots,y_K)}, \tag{1}$$

where $\theta^{(0)}$ represents the regression coefficient corresponding to a referent subtype of the disease, the coefficients $\theta^{(1)}_{k(y_k)}$ represent the first-order parameter contrasts, $\theta^{(2)}_{kk'(y_k,y'_k)}$ denote the second-order parameter contrasts, and so on.

The above representation of the hazard ratio parameters suggests a natural and hierarchical way of reducing the number of parameters by constraining suitable contrasts to be zero. For instance, if we assume that the second and all higher-order contrasts are equal to zero, then (1) reduces to

$$\beta_{(y_1,\cdots,y_K)} = \theta^{(0)} + \sum_{k=1}^{K} \theta^{(1)}_{k(y_k)}, \tag{2}$$

and in this case the heterogeneity between two subtype-specific regression coefficients $\beta_{(y_1,\cdots,y_{k-1},y_k,y_{k+1},\cdots,y_K)} - \beta_{(y_1,\cdots,y_{k-1},y^*_k,y_{k+1},\cdots,y_K)} = \theta^{(1)}_{k(y_k)} - \theta^{(1)}_{k(y^*_k)}$. Thus, in this model, the contrasts of the form $\theta^{(1)}_{k(y_k)} - \theta^{(1)}_{k(y^*_k)}$ give a measure of the degree of etiologic heterogeneity among the disease subtypes with respect to the k-th trait, holding the level of other traits to be constant. Further, for ordinal or continuous disease traits, one can impose ordering constraints on the contrast parameters by using models of the form

$$\theta^{(1)}_{k(y_k)} = \theta^{(1)}_k s^{(k)}_{y_k}, \quad y_k = 1, \ldots, M_k, \tag{3}$$

where $\left\{0 = s_1^{(k)} \leq s_2^{(k)} \ldots \leq s_{M_k}^{(k)}\right\}$ are a set of pre-specified scores assigned to the M_k levels of the k-th trait. This model summarizes the degree of etiologic heterogeneity with respect to the k-th characteristic in terms of a single regression coefficient $\theta_k^{(1)}$, with $\theta_k^{(1)} = 0$ implying no heterogeneity.

The additive model (2) could be extended further to include interaction terms between pairs of disease traits, thus potentially allowing the degree of etiologic heterogeneity with respect to one trait to vary with the level of the other trait, and vice versa. For ordinal traits, the interaction parameters can also be constrained to maintain their ordering by modeling them in terms of underlying continuous scores. More details about these modeling techniques can be found in Chatterjee (2004).

The second-stage model described above can be represented as $\beta = \mathcal{B}\theta$, where the design matrix $\mathcal{B}$ relates the coefficients $\boldsymbol{\beta}$ of the unstructured competing-risk Cox model to the parameters θ of the log-linear model. When there are multiple covariates of interest, say $X = (X_1, \ldots, X_P)$, then one can consider a separate two-stage model of the form $\beta_p = \mathcal{B}^{(p)}\theta_p$ for each of the covariates X_p. In the following exposition, we will assume $\mathcal{B}^{(p)} = \mathcal{B}$ for all p for notational convenience.

3. Inference

3.1. *Estimation Methodology*

If there are no missing data on any of the disease traits, inference on the covariate regression parameters can be performed based on a partial likelihood of the form

$$L_n = \prod_{i:\Delta_i=1} \left\{ \frac{h_{Y_i}(V_i|X_i)}{\sum_{j=1}^n I(V_j \geq V_i) h_{Y_i}(V_i|X_j)} \right\} = \prod_{i:\Delta_i=1} \left\{ \frac{\exp(\sum_p \mathcal{B}_{Y_i}^T \theta_p X_{ip})}{\sum_{j=1}^n I(V_j \geq V_i) \exp(\sum_p \mathcal{B}_{Y_i}^T \theta_p X_{jp})} \right\}, \quad (4)$$

where $Y_i = (Y_{i1}, \ldots, Y_{iK})$ denotes the observed disease traits for the i-th subject given $\Delta_i = 1$ (case) and $\mathcal{B}_{Y_i}^T$ represents the row of the design matrix $\mathcal{B}$ that corresponds to the trait values Y_i. The maximum partial likelihood estimate of θ can be obtained by

maximizing (4) without requiring any assumption about the subtype-specific baseline hazard functions $\lambda_y(t)$. The score function associated with (4) can be written in the general form

$$S^*_\theta \;=\; \sum_{\Delta_i=1} \left\{ \mathcal{B}_{Y_i} \otimes X_i - \frac{\sum_{j=1}^n I(V_j \geq V_i) \mathcal{B}_{Y_i} \otimes X_j \exp(\sum_p \mathcal{B}^T_{Y_i} \theta_p X_{jp})}{\sum_{j=1}^n I(V_j \geq V_i) \exp(\sum_p \mathcal{B}^T_{Y_i} \theta_p X_{jp})} \right\}, \qquad (5)$$

where $\otimes$ denotes the Kronecker product. If data on some or all of the disease traits are missing for some study subjects, one can use (4) to perform a complete-case analysis involving only those cases who have complete data on all the disease traits. Such complete-case analyses, however, potentially may require discarding large numbers of cases resulting in inevitable loss of efficiency and possibly bias due to non-random missingness.

Several researchers in the past have studied the problem of missing cause-of-failure with two underlying competing events. Dewanji (1992) proposed a partial-likelihood-based solution by assuming that the baseline hazards for the two causes of failure are proportional. Goetghebeur & Ryan (1995) also exploited the same assumption but proposed a more robust estimation equation approach that is less sensitive to misspecification of the constant hazard-ratio assumption. Craiu & Duchesne (2004) proposed a full-likelihood-based solution allowing for flexible piecewise-constant modeling of the baseline hazard functions. Lu & Tsiastis (2001) considered a multiple-imputation method that estimates the odds of one failure type against the other as a function of failure time and covariates by parametric logistic regression analysis of the complete-case data. Gao & Tsiatis (2005) proposed an inverse probability-weighted complete case estimator and a doubly robust estimator in general linear transformation models. Although each of these methods has its own advantages, we find that the method of Goetghebeur & Ryan (1995) is particularly appealing for extension to our setting, as this method is quite efficient and yet relies minimally on the assumptions about the failure-type-specific baseline hazard functions. In particular, when there are no missing data on cause of failure, the method boils down to standard partial likelihood method. When there are missing data,

although some assumptions about the interrelationship among the event-specific baseline hazard functions are needed, the method remains relatively robust, as it relies on those assumptions only when it is needed for dealing with events with missing cause of failure. In the following, we propose a general extension of the approach of Goetghebeur & Ryan (1995) for a structured competing-risk problem where the competing events are defined by underlying related disease traits.

We first introduce some additional notations. Let R_k denote the indicator of whether the k-th disease trait is observed ($R_k = 1$) or not ($R_k = 0$), for a diseased subject. Define $R = (R_1, \ldots, R_K)$ and let $r = (r_1, \ldots, r_K)$ be a realization of R. We observe that r can take on 2^K possible values, each corresponding to a particular pattern of missing data for the disease traits. We will assume that the data are missing-at-random, in the sense that

$$\Pr(R = r|T, X, Y, \Delta = 1) = \Pr(R = r|T, X, \Delta = 1) = \pi^{(r)}(T, X), \tag{6}$$

i.e., the probabilities for the different patterns of missing data do not depend on the trait values Y. The model (6) is more general than that considered by Goetghebeur & Ryan, as they allowed the missingness probability to depend only on T, but not on X. For any given patten of missingness r, partition the disease traits as $Y = (Y^{o_r}, Y^{m_r})$, where Y^{o_r} and Y^{m_r} denote the vectors of traits that have been observed and missing, respectively. Now, we can write the hazard of the disease with a missingness pattern $R = r$ and disease trait y^{o_r} as

$$h_{y^{o_r}}(t|X) \;=\; \int_{y^{m_r}} \lambda_{(y^{m_r}, y^{o_r})}(t) \exp\{\beta_{(y^{m_r}, y^{o_r})} X_i\} d\mu(y^{m_r}),$$

where $\mu(\cdot)$ denotes a suitable measure on the sample space of y^{m_r}.

Define

$$Q_{y^{o_r}}^{y^{m_r}}(t, X) = \frac{h_{(y^{m_r}, y^{o_r})}(t|X)}{h_{y^{o_r}}(t|X)}$$

and observe that $Q_{y^{o_r}}^{y^{m_r}}(t, X)$ can be interpreted as the conditional probability that a diseased subject with the incidence time t and the covariate value X has trait value

$y = (y^{m_r}, y^{o_r})$, given the observed disease traits y^{o_r}. Now define

$$\mathcal{E}_{y^{o_r}}(\mathcal{B}_Y|T=t,X) = \int_{y^{m_r}} \mathcal{B}_{(y^{o_r},y^{m_r})} Q_{y^{o_r}}^{y^{m_r}}(t,X) d\mu(y^{m_r})$$

to be the conditional expectation of the design vector $\mathcal{B}_Y$ given the observed traits y^{o_r}, the time to first disease occurrence $T = t$, and the covariate data X. Now, we propose to estimate θ based on the estimating function

$$\begin{aligned} S_\theta \quad &\equiv \sum_r S_\theta^{(r)} \equiv \sum_r \sum_{\Delta_i=1, R_i=r} \Big\{ \mathcal{E}_{y_i^{o_r}}(\mathcal{B}_Y|T_i, X_i) \otimes X_i \\ &- \frac{\sum_{j=1}^n I(V_j \geq V_i) \mathcal{E}_{y_i^{o_r}}(\mathcal{B}_Y|T_i, X_j) \otimes X_j h_{y_i^{o_r}}(T_i|X_j)}{\sum_{j=1}^n I(V_j \geq V_i) h_{y_i^{o_r}}(T_i|X_j).} \Big\} = 0. \end{aligned} \tag{7}$$

The unbiasedness of the estimating equation (7) under the general missing-at-random model (6) is proved in the Appendix. In the absence of any missing data on disease traits, $\mathcal{E}_{y_i^{o_r}}(\mathcal{B}_Y|X)$ corresponds to the row of the second-stage design matrix $\mathcal{B}$ associated with the observed disease traits y_i and the estimating equation (7) is equivalent to the score equation associated with the partial likelihood (4). If we assumed discrete disease subtypes and did not impose any structure through the second-stage model, then $\mathcal{E}_{y_i^o}(\mathcal{B}_Y|X_i, T_i)$ would simply correspond to the conditional probability vector for observing the different disease subtypes, given the observed traits, and the estimating equation (7) would essentially very similar to that proposed by Goetghebeur and Ryan for dealing with two failure types.

The estimating equation (7) cannot be used by itself, as it involves the "nuisance" baseline hazard function $\lambda_y(t)$. A complete nonparametric estimation of $\lambda_y(t)$ may not be feasible because when the number of disease subtypes gets large, the number of cases for the individual subtypes gets sparse. In the following, we consider a semi-parametric estimation approach for the baseline hazard functions. First, we express

$$\lambda_y(t) = \lambda_{(1,\cdots,1)}(t) \exp\{\alpha_y(t)\},$$

where $\lambda_{(1,\cdots,1)}(t)$ denotes the baseline hazard associated with a reference disease subtype and $\exp\{\alpha_y(t)\}$ denotes the baseline hazard for the disease subtype y expressed as a

multiple of that for the reference subtype. Similar to Goetghebeur and Ryan, we now invoke an assumption of a time-independent hazard-ratio, that is, $\alpha_y(t) \equiv \alpha_y$ for all values of y. In addition, to overcome the potential sparsity problem when there are a large number of disease subtypes, we propose to specify $\exp(\alpha_y)$ using log-linear models analogous to those we used to specify the covariate hazard-ratio parameters $\exp(\beta_y)$. We could, for example, consider an additive model of the form

$$\alpha_{(y_1,\ldots,y_K)} = \xi^{(0)} + \sum_{k=1}^{K} \xi^{(1)}_{k(y_k)}.$$

Let $\alpha = \mathcal{A}\xi$ represent such a model, and let $\mathcal{A}_Y$ denote the row of the design matrix $\mathcal{A}$ which corresponds to the trait Y. Define

$$h_u(t|X) = \int_y \lambda_y(t_i) \exp(\beta_y X_i) d\mu(y),$$

$$Q^y_u(X) = \frac{\exp(\alpha_y + \beta_y X)}{\int_y \exp(\alpha_y + \beta_y X_i) d\mu(y)}, \quad w_u(X) = \int_y \exp(\alpha_y + \beta_y X) d\mu(y)$$

and observe that $h_u(t|X)$ denotes the marginal hazard for the disease integrated over all different values of the disease traits and $Q^y_u(X_i)$ is the probability that a subject has disease trait y, given simply that it is a case ($\Delta = 1$) but ignoring all other disease trait information.

Under the constant-hazard-ratio assumption $\alpha_y(t) \equiv \alpha_y$, the quantities $Q^{y^{m_r}}_{y^{o_r}}(t, X)$ and $h_{y^{o_r}}(t|X)$ in the definition of estimating equation (7) can be replaced by

$$Q^{y^{m_r}}_{y^{o_r}}(X) = \frac{\exp\{\alpha_{(y^{m_r}, y^{o_r})} + \beta_{(y^{m_r}, y^{o_r})} X\}}{\int_{y^{m_r}} \exp\{\alpha_{(y^{o_r}, y^{m_r})} + \beta_{(y^{o_r}, y^{m_r})} X\} d\mu(y^{m_r})}$$

and

$$w_{y^{o_r}}(X) = \int_{y^{m_r}} \exp\{\alpha_{(y^{o_r}, y^{m_r})} + \beta_{(y^{o_r}, y^{m_r})} X\} d\mu(y^{m_r}),$$

respectively. We further let

$$\mathcal{E}_{y^{o_r}}(\mathcal{A}_Y|X) = \int_{y^{m_r}} \mathcal{A}_{(y^{o_r}, y^{m_r})} Q^{y^{m_r}}_{y^{o_r}}(X) d\mu(y^{m_r}) \text{ and } \mathcal{E}_u(\mathcal{A}_Y|X) = \int_y \mathcal{A}_y Q^y_u(X) d\mu(y).$$

Now, following Gotnberger and Ryan, we propose to estimate ξ based on the partial-likelihood

$$L_n^* = \prod_r \prod_{R_i=r,\Delta_i=1} \Big\{ \frac{h_{y_i^{or}}(V_i|X_i)}{\sum_{j=1}^n I(V_j \geq V_i) h_u(V_i|X_j)} \Big\}.$$

The associated score function can be conveniently expressed as

$$S_\xi \equiv \sum_r S_\xi^{(r)} \equiv \sum_r \sum_{\Delta_i=1,R_i=r} \Big\{ \mathcal{E}_{y_i^{or}}(\mathcal{A}_Y|X_i) - \frac{\sum_{j=1}^n I(V_j \geq V_i) \mathcal{E}_u(\mathcal{A}_Y|X_j)\omega_u(X_j)}{\sum_{j=1}^n I(V_j \geq V_i)\omega_u(X_j)} \Big\}. \quad (8)$$

In our applications, we solved both sets of estimating equations $S_\theta = 0$ and $S_\xi = 0$ by using Newton-Raphson method. In principle, these equations can also be solved by EM-type algorithm where the expectation steps will involve computing the conditional expectations $\mathcal{E}_{y_i^O}(\mathcal{B}_Y|X_i, T_i)$ and $\mathcal{E}_{y_i^{or}}(\mathcal{A}_Y|X_i)$, and then the M-step will involve solving partial-likelihood-like score equations of the form (7).

3.2. *Asymptotic Theory and Variance Estimation*

In this section, we study the asymptotic theory for the proposed estimator. Unlike a martingale-based approach considered by Goetghebeur and Ryan, here we consider an empirical process representation of the score functions to derive the influence function of the proposed estimator and an associated robust sandwich estimator for the asymptotic variance. Application of the robust sandwich estimator, as opposed to a model based estimator, in this setting is particularly appealing because of the possibility of the mis-specification of the models for the baseline hazard functions as a function of time t and the disease trait y. In the following lemma, we first provide a result about asymptotic unbiasedness of the estimating functions.

LEMMA 1. *Under the regularity conditions listed in the Appendix and the missing-at -random mechanism (6), as $n \to \infty$,*

$$\frac{1}{n} S_\theta \equiv \frac{1}{n} \sum_r S_\theta^{(r)} \to_p 0 \quad \textit{and} \quad \frac{1}{n} S_\xi \equiv \frac{1}{n} \sum_r S_\xi^{(r)} \to_p 0,$$

at the true parameter values. Moreover, if $\pi(T, X) = \pi(T)$,

$$\frac{1}{n} S_\theta^{(r)} \to_p 0 \quad \textit{and} \quad \frac{1}{n} S_\xi^{(r)} \to_p 0 \quad \textit{for each specific} \quad r.$$

An interesting corollary of the above lemma is that the complete-case analysis, which corresponds to $r = (1, \ldots, 1)$, is unbiased when the missingness probability depends only on T, but not on X. Lemma 1 also shows that the martingale representation that Goetghebeur & Ryan used for their estimating function for each specific type of missing data pattern is not valid under the more general missing data model we consider. Now, we define some further notation. Let

$$S_\theta^{(1)}(V_i, Y_i^{o_r}) = \frac{1}{n} \sum_{j=1}^{n} I(V_j \geq V_i) \mathcal{E}_{Y_i^{o_r}}(\mathcal{B}_Y | X_j) \otimes X_j \omega_{Y_i^{o_r}}(X_j),$$

$$S_\xi^{(1)}(V_i, u) = \frac{1}{n} \sum_{j=1}^{n} I(V_j \geq V_i) \mathcal{E}_u(\mathcal{A}_Y | X_j) \omega_u(X_j),$$

$$S^{(0)}(V_i, Y_i^{o_r}) = \frac{1}{n} \sum_{j=1}^{n} I(V_j \geq V_i) \omega_{Y_i^{o_r}}(X_j), \;\; S^{(0)}(V_i, u) = \frac{1}{n} \sum_{j=1}^{n} I(V_j \geq V_i) \omega_u(X_j),$$

and denote by $s^{(1)}(V_i, Y_i^{o_r})$, $s^{(0)}(V_i, Y_i^{o_r})$, $s^{(1)}(V_i, u)$, and $s^{(0)}(V_i, u)$ the corresponding population expectations. The parameter vector $\eta = (\theta^\top, \xi^\top)^T$ is estimated by solving

$$T_n = \begin{pmatrix} S_\theta \\ S_\xi \end{pmatrix} = 0. \tag{9}$$

Define $\mathcal{I} = \lim_{n \to \infty} (1/n) \partial T_n / \partial \eta$. The formulae for the various components of $\mathcal{I}$ are given in the Appendix. The following theorem summarizes the asymptotic property of the estimators.

THEOREM 1. *Under the regularity conditions listed in the Appendix, the estimating equation* $T_n = 0$ *has a unique consistent sequence of solutions* $(\hat{\theta}_n, \hat{\xi}_n)$ *that is asymptotically normally distributed, with the influence function representation given by*

$$n^{1/2} \begin{pmatrix} \hat{\theta}_n - \theta_0 \\ \hat{\xi}_n - \xi_0 \end{pmatrix} = \mathcal{I}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} J_{ni}^\theta \\ J_{ni}^\xi \end{pmatrix} + o_p(1), \tag{10}$$

where

$$J_{ni}^{\theta} = \Delta_i \sum_r I(R_i = r)\Big\{\mathcal{E}_{Y_i^{o_r}}(\mathcal{B}_Y|X_i)\otimes X_i - \frac{s^{(1)}(V_i, Y_i^{o_r})}{s^{(0)}(V_i, Y_i^{o_r})}\Big\} + \sum_r E_{\Delta,R,V,Y^{o_r}}\left[\Delta I(R=r)\frac{I(V\le V_i)\omega_{Y^{o_r}}(X_i)}{s^{(0)}(V,Y^{o_r})}\left\{\mathcal{E}_{Y^{o_r}}(\mathcal{B}_Y|X_i)\otimes X_i - \frac{s^{(1)}(V,Y^{o_r})}{s^{(0)}(V,Y^{o_r})}\right\}\right]$$

and

$$J_{ni}^{\xi} = \Delta_i \sum_r I(R_i = r)\Big\{\mathcal{E}_{Y_i^{o_r}}(\mathcal{A}_Y|X_i)\otimes X_i - \frac{s^{(1)}(V_i, u)}{s^{(0)}(V_i, u)}\Big\} + E_{\Delta,V}\left[\Delta\frac{I(V\le V_i)\omega_u(X_i)}{s^{(0)}(V,u)}\left\{\mathcal{E}_u(\mathcal{A}_Y|X_i) - \frac{s^{(1)}(V,u)}{s^{(0)}(V,u)}\right\}\right].$$

To obtain an "empirical" sandwich variance estimator, we can estimate J_{ni}^{θ} and J_{ni}^{ξ} by

$$\hat{J}_{ni}^{\theta} = \Delta_i \sum_r I(R_i = r)\Big\{\mathcal{E}_{Y_i^{o_r}}(\mathcal{B}_Y|X_i)\otimes X_i - \frac{S^{(1)}(V_i, Y_i^{o_r})}{S^{(0)}(V_i, Y_i^{o_r})}\Big\} + \sum_r\sum_{l=1}^{n}\Delta_l I(R_l = r)\frac{I(V_l\le V_i)\omega_{Y_l^{o_r}}(X_i)}{S^{(0)}(V_l, Y_l^{o_r})}\left\{\mathcal{E}_{Y_l^{o_r}}(\mathcal{B}_Y|X_i)\otimes X_i - \frac{S^{(1)}(V_l, Y_l^{o_r})}{S^{(0)}(V_l, Y_l^{o_r})}\right\}$$

and

$$\hat{J}_{ni}^{\xi} = \Delta_i \sum_r I(R_i = r)\Big\{\mathcal{E}_{Y_i^{o_r}}(\mathcal{A}_Y|X_i)\otimes X_i - \frac{S^{(1)}(V_i, u)}{S^{(0)}(V_i, u)}\Big\} + \sum_{l=1}^{n}\Delta_l\frac{I(V_l\le V_i)\omega_u(X_i)}{S^{(0)}(V_l,u)}\left\{\mathcal{E}_u(\mathcal{A}_Y|X_i) - \frac{S^{(1)}(V_l,u)}{S^{(0)}(V_l,u)}\right\}$$

The sandwich variance estimator for $\hat{\eta} = (\hat{\theta}^\top, \hat{\xi}^\top)^\top$ can now be obtained as $\hat{\mathcal{I}}^{-1}\hat{V}\hat{\mathcal{I}}^{-T}$, where $\hat{\mathcal{I}} = \partial T_n/\partial\eta$ (see the supplementary materials for the formula) and $\hat{V} = \sum_{i=1}^{n}\hat{J}_{ni}\hat{J}_{ni}^{T}$.

4. Simulation Study

In this section, we study the finite sample performance of the proposed estimator under correct and misspecified models for the baseline hazard functions. We consider two

different simulation scenarios. In the first, we assume the total number of distinct disease traits to be small and the number of cases observed for each trait value to be reasonably large. Within this first scenario, we consider trait values to be missing-completely-at-random and missing-at-random. In the second setting, we allow the number of distinct disease traits to be large and the number of cases observed for the different trait values to be potentially sparse.

4.1. *Moderate number of disease subtypes*

In this setting, we simulated data by mimicking the incidence pattern of four subtypes of breast cancer: ER+PR+, ER+PR-, ER-PR+, and ER-PR-. We denote the four disease subtypes as (2, 2), (2, 1), (1, 2), and (1, 1). We first generated a scalar covariate X from the Normal$(0, \sigma = 1)$ distribution for a cohort of size n, with $n = 10,000$. Next we generated the age-at-onset for the four different subtypes of the disease using the proportional hazards model

$$\lambda_{(y_1,y_2)}(t|X) = \lambda_{(y_1,y_2)}(t) \exp(\beta_{(y_1,y_2)}X),$$

with a Weibull specification for the baseline hazard functions,

$$\lambda_{(y_1,y_2)}(t) = \gamma_{(y_1,y_2)}\lambda_{(y_1,y_2)}^{\gamma_{(y_1,y_2)}} t^{\gamma_{(y_1,y_2)}-1}, \tag{11}$$

for $y_1 = 1, 2$ and $y_2 = 1, 2$.

We assumed that $\beta_{(y_1,y_2)} = \theta^{(0)} + \theta^{(1)}_{1(y_1)} + \theta^{(1)}_{2(y_2)}$ with $\theta^{(1)}_{1(1)} = \theta^{(1)}_{2(1)} = 0$ for identifiability. We set the value of $\theta^{(0)}$, a common parameter across all subtypes, to be 0.35, which indicated an overall positive association between the covariate and the hazard of the disease. In addition, we set $\theta^{(1)}_{2(2)} = 0.25$, implying a stronger effect of the covariate on the risk of PR+ tumors compared with PR- tumors. We assumed $\theta^{(1)}_{1(2)} = 0$, that is, no difference in the effect of the covariate between ER+ and ER- tumors. We assume $\gamma_{y_1,y_2} = \gamma = 4.355$ for all (y_1, y_2), which guaranteed constant hazard ratios between the

different subtypes. We next further specifies $\lambda_{(y_1,y_2)}$ in such a way that the hazard ratios $\alpha_{(y_1,y_2)} = \{\lambda_{(y_1,y_2)}/\lambda_{(1,1)}\}^{\gamma}$ satisfied the constraint

$$\log\{\alpha_{(y_1,y_2)}\} = \xi^{(0)} + \xi^{(1)}_{1(y_1)} + \xi^{(1)}_{2(y_2)}, \tag{12}$$

with $\xi^{(0)} = 0$, $\xi^{(1)}_{1(1)} = \xi^{(1)}_{2(1)} = 0$, $\xi^{(1)}_{1(2)} = -0.0295$ and $\xi^{(1)}_{2(2)} = -2.1779$.

We generated the censoring time C for the subjects using a Normal$(75, 5^2)$ distribution. In our simulation setting, we observed, on average, 3.912%, 3.864%, 0.562%, and 0.514% cases of subtypes $(2,2), (2,1), (1,2)$, and $(1,1)$, respectively. After generating data for the entire cohort, we simulated missing trait values under missing-completely-at-random and missing-at-random models. Under missing-completely-at-random, we deleted ER and PR status randomly, independently of each other, using Bernoulli sampling with the probability of missing ($p = 1 - \pi$) for each being 0.20 or 0.50. Under missing-at-random, we assumed the trait values to be completely observed for cases whose X and T values are greater than the 80^{th} percentile values for the respective distributions. Among the remaining cases, we simulated missing traits by deleting ER and PR status randomly, independently of each other, using Bernoulli sampling with the probability of missing data (p) for each being 0.31 and 0.78, so that the overall missingness probability for each trait is, approximately, 0.20 or 0.50 respectively.

Under each scenario, we simulated 500 data sets and analyzed each of them using three different methods. In the first, subsequently referred to as *Full Cohort* analysis, we applied the partial likelihood (4) to the entire cohort, assuming that data on both of the disease traits had been observed for all the cases. In the second, subsequently referred to as *Complete-Case* analysis, we applied the partial likelihood (4) to the cohort after deleting all the cases who had missing data in any of the traits. In the third, subsequently referred to as the *Estimating-Equation* analysis, we applied the proposed method to analyze data on all the cases, including those with one or both of the two traits missing.

The results presented in Table 1 reveal that the proposed estimating equation

method had negligible bias for estimation of both the null and non-null values of the θ-parameters. Moreover, the proposed robust sandwich variance estimator also performed very well for estimation of the true variance of the parameter estimates. The coverage probabilities for the associated 95% confidence intervals were also close to the nominal level. Under missing-completely-at-random, the complete-case analysis also produced unbiased parameter estimates. The estimating equation method, however, was clearly more efficient, as it incorporated data on all cases irrespective of whether they had missing traits or not. In fact, when the proportion of missing was modest, e.g 20%, the estimating equation method lost only modest efficiency - in the range of 4-15% - compared with the full-cohort analysis. In contrast, the loss of efficiency for complete-case analysis ranged between 30 and 40%, even with a modest amount of missing data. Under missing-at-random, the complete-case analysis yielded a biased estimate of the parameter $\theta^{(0)}$, with the magnitude of bias being quite high when the missingness probability was 0.50. Moreover, under missing-at-random, the loss of efficiency for the complete case analysis was more dramatic than that observed under missing-completely-at-random.

Next we investigated the robustness of the estimating-equation method against misspecification of the semiparametric model for the trait specific baseline hazard functions. The simulation design remained the same as above, except that now in model (11) we allowed the parameters $\gamma_{(y_1,y_2)}$ and $\lambda_{(y_1,y_2)}$ to vary freely. In particular, we set $\gamma_{(1,1)} = 2.692$, $\gamma_{(1,2)} = 4.058$, $\gamma_{(2,1)} = 5.433$, $\gamma_{(2,2)} = 4.355$, and $\lambda_{(1,1)} = 0.0026$, $\lambda_{(1,2)} = 0.0038$, $\lambda_{(2,1)} = 0.0063$ and $\lambda_{(2,2)} = 0.0072$, giving rise to approximately 1.34%, 1.64%, 0.73% and 7.06% disease incidence of subtypes (1,1), (2,1), (1,2), and (2,2), respectively, in the underlying cohort. Figure 1 shows how $\log\{\lambda_{(y_1,y_2)}(t)/\lambda_{(1,1)}(t)\}$ changes over time as opposed to taking a constant value under the "working" model for the estimating-equation method. **From the results shown in Table 2, we observe that in this setting, under missing-completely-at-random all of the methods produced nearly unbiased estimates for all the parameters, but some noticeable bias**

was observed for the estimating-equation method for the estimation of the parameter $\theta^{(1)}_{1(2)}$ under the setting of 50% missing data. Under missing-at-random, in contrast, the complete-case analysis produced severe bias in estimating the parameter $\theta^{(0)}$ and the corresponding 95% coverage probability was unacceptably low. Under missing-at-random, the bias of the estimating-equation method also increased, but still remained small in absolute terms and the corresponding 95% coverage probabilities were reasonable.

4.2. *A Large number of disease subtypes*

In the third and final setting, we considered three disease traits, each with four levels, say $y_j = 1, 2, 3$, and 4, with the total number of disease subtypes equal to $4 \times 4 \times 4 = 64$. As earlier, we generated the failure times for different disease subtypes from a trait-specific Cox proportional hazards model, where the covariate log-hazard-ratio parameters satisfied the constraint

$$\beta_{(y_1,y_2,y_3)} = \theta^{(0)} + \theta^{(1)}_1 s_{1(y_1)} + \theta^{(1)}_2 s_{2(y_2)} + \theta^{(1)}_3 s_{3(y_3)},$$

with $s_{j(y_j)} = (y_j - 1)^{0.3}$ (see Chatterjee, 2004) and the baseline hazard functions followed Weibull distribution of the form (11). We chose $\theta^{(0)} = 0.35$, $\theta^{(1)}_1 = 0.15$, $\theta^{(1)}_2 = 0.0$ and $\theta^{(1)}_3 = 0.5$. We allowed 64 unrestricted values for the λ- and γ- parameters of the baseline Weibull distribution by randomly drawing their values from the Uniform$(3.5, 4)$ and Uniform$(0.0021, 0.0024)$ distributions, respectively. As before, $X \sim$ Normal$(0, 1)$ and the censoring time was generated from a Normal$(75, 5^2)$ distribution. In this setting, the fraction of the subjects in the cohort who developed the disease was approximately 11%, with the subtype-specific disease occurrence rates ranging between 0.076% and 0.314%. We considered two different sample sizes, $n = 5,000$ and $10,000$.

As before, we analyzed each data set using three methods: full-cohort, complete-case and estimating-equation. For the estimating-equation method, we assumed constant

hazard ratios across subtypes and a working model of the form $\log\{\alpha_{(y_1,y_2,y_3)}\} = \xi^{(0)} + \xi^{(1)}_{1(y_1)} + \xi^{(2)}_{2(y_2)} + \xi^{(3)}_{3(y_3)}$. The results from this simulation study (shown in Table 1 of supplemental materials) reveal that all of the different methods produced valid inferences in this setting of highly stratified disease subtypes. The bias of the estimating-equation method, along with that for the complete-case and the full-cohort analyses, was small even though the working model for the baseline hazard functions was incorrectly specified for the first method. Further, the estimating-equation method often gained remarkable efficiency compared with the complete-case analysis.

5. Analysis of the Cancer Prevention Study II Nutrition Cohort

The Cancer Prevention Study (CPS)-II Nutrition Cohort is a prospective study of cancer incidence and mortality among men and women in the United States that was established in 1992 and was ended on June 30, 2005.

In brief, the study participants completed a mailed, self-administered questionnaire in 1992 or 1993 that included a food frequency diet assessment and information on demographic, medical, behavioral, environmental, and occupational factors. Beginning in 1997, follow-up questionnaires were sent to cohort members every 2 years to update exposure information and to ascertain newly diagnosed cancers; response rates for all followup questionnaires have been at least 90% (for details see Feigelson et al., 2006). For the purpose of illustration, we considered weight gain from age 18 to the year 1992 as the main covariate of interest as it has previously been shown to be related to risk of breast cancer in the CPS-II cohort. After excluding women who were either lost to follow-up, had unknown weights, had extreme values of weight, or reported prevalent breast or other cancer at baseline, except nonmelanoma skin cancer, we were left with $44,172$ women who are postmenopausal at baseline in 1992.

Among the 44,172 women, we found that 1516 had some form of breast cancer. The cancer cases were verified by obtaining medical records or through linkage with state cancer registries when complete medical records could not be obtained. We analyze available data on five tumor traits; (1) *Grade*, with three categories: well/moderately/poorly differentiated; (2) *Stage*, with two categories: localized/distant; (3) *Histologic type*, with three categories: ductal/lobular/other; (4) *Estrogen receptor (ER) status* with two categories: ER+/ER-; (5) *Progesterone receptor (PR) status*, with two categories PR+/PR-. The aim of our analysis was to study how the association between weight gain and risk of breast cancer varied by various tumor traits.

The five traits yielded a total of up to $3 \times 2 \times 3 \times 2 \times 2 = 72$ subtypes. Out of the 1516 cancer patients, 782 subjects were information on all of the disease traits, while the remaining 734 subjects had information on at least one of the traits missing. Let y_1, y_2, y_3, y_4, and y_5 denote the level of grade, stage, histology, ER status, and PR status, respectively. We modeled the hazard of the various cancer subtypes as

$$h_{(y_1,y_2,y_3,y_4,y_5)}(t|X) = \lambda_{(y_1,y_2,y_3,y_4,y_5)}(t) \exp\{X\beta_{(y_1,y_2,y_3,y_4,y_5)}\}, \tag{13}$$

where, further, we assumed $\beta_{(y_1,y_2,y_3,y_4,y_5)} = \theta^{(0)} + \theta^{(1)}_{1(y_1)} + \theta^{(1)}_{2(y_2)} + \theta^{(1)}_{3(y_3)} + \theta^{(1)}_{4(y_4)} + \theta^{(1)}_{5(y_5)}$. We set well-differentiated, localized, ductal, ER+ and PR+ as the referent levels for the associated tumor traits. Thus, in our model, $\theta^{(0)}$ represents the log hazard ratio of weight gain associated with this referent breast cancer subtype and the parameters $\theta^{(1)}_{k(y_k)}$, $k = 1, 2, 3, 4, 5$, yield a measure of heterogeneity in the effect of weight gain by the levels of the corresponding disease trait. We performed both the complete-case and the estimating-equation analysis of the data. For our estimating-equation method, we assumed the constraint

$$\log\{\lambda_{(y_1,y_2,y_3,y_4,y_5)}(t)/\lambda_{(1,1,1,1,1)}(t)\} = \xi^{(1)}_{1(y_1)} + \xi^{(1)}_{2(y_2)} + \xi^{(1)}_{3(y_3)} + \xi^{(1)}_{4(y_4)} + \xi^{(1)}_{5(y_5)}.$$

From the results presented in Table 3, it is evident that both methods produced estimates of $\theta^{(0)}$ to be positive and highly significant, indicating an overall positive association of

weight gain with the risk of breast cancer. Moreover, the significance of the estimate of $\theta^{(1)}_{2(2)}$ in both the methods indicated that the association between weight gain and risk of breast cancer is stronger for distant compared with localized tumors. The significance of the estimate of $\theta^{(1)}_{5(2)}$ in both methods also suggests that the association between weight gain and risk of breast cancer is stronger for PR+ compared with PR- tumors. For all parameters, the estimating-equation method produces substantially smaller standard errors compared with the complete-case analysis.

6. Discussion

In this article, we have considered an estimating-equation approach for inference in the proposed two-stage proportional hazards regression model in the presence of missing disease trait information. The proposed method is semiparametric in the sense that it involves an unspecified baseline hazard function for a baseline disease subtype. The method requires an assumption of missing-at-random but it does not require any modeling assumption for $\pi(T, X)$, the missingness probability, for the disease traits. The asymptotic unbiasedness of the method, however, does require correct parametric specification for the interrelationships of the nuisance baseline hazard functions for the different disease subtypes.

Our extensive simulation study suggests that in practice, the magnitude of bias generated by the proposed method for estimation of θ, the focus parameters of interest, is generally quite small, even when the model for $\lambda_y(t)/\lambda_{(1,\ldots,1)}(t)$ is grossly misspecified. The theory we have developed is valid for more general parametric specification of $\lambda_y(t)/\lambda_{(1,\ldots,1)}(t)$ which does not necessarily assume constancy over t. Thus the proposed methods can be used to conduct sensitivity analyses against alternative models for the baseline hazards. In principle, one could also construct a doubly robust estimator for θ along the lines of Gao & Tsiastis (2005). In particular, one could use the constant hazard ratio and log-linear modeling assumption for the baseline hazard functions to

develop a doubly-robust estimator of θ that would be unbiased for a correctly specified model for $\pi(T, X)$ or for $\lambda_y(t)/\lambda_{(1,\ldots,1)}(t)$. Hence, it would be interesting to develop such a method and to examine how its gain in robustness compensates for its potential loss of efficiency compared with the estimating equation method in practical settings.

References

Chatterjee, N. (2004). A two-stage regression model for epidemiological studies with multivariate disease classification data. *Journal of the American Statistical Association*, **99**, 127–138.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Craiu, R. V. & Duchesne, T. (2004). Inference based on the EM algorithm for the competing risks model with masked causes of failure. *Biometrika*, **91**, 543–558.

Dewanji, A. (1992). A note on a test for competing risks with missing failure type. *Biometrika*, **79**, 855–857.

Feigelson, H. S., Patel, A. V., Teras, L. R., Gansler, T., Thusn, M. J., & Calle, E. E. (2006). Adult weight gain and histopathologic characteristics of breast cancer among postmenopausal women. *Cancer*, **107**, 12–21.

Gao, G. & Tsiatis, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika*, **92**, 875–891.

Goetghebeur, E. & Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika*, **82**, 821–833.

Lu, K. & Tsiatis, A. A. (2005). Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. *Lifetime Data*

Analysis, **11**, 29–40.

Vaart, A. W. van der. (1998). *Asymptotic Statistics*, Cambridge University Press: Cambridge, UK.

Appendix

Regularity Conditions

In the following, let $|\theta|_p$ denote the sum of the absolute values of the θ-parameters associated with the vector of covariates X_p.

(A1) $\pi^{(r)}(T,S) > 0$ almost surely in (T,S) for $r = (1,1,\ldots,1)$.

(A2) The elements of the second-stage design matrices $\mathcal{B}$ and $\mathcal{A}$ remain uniformly bounded in absolute value by constants, say c_B and c_A, respectively.

(A3) Assume the function $X^{\otimes 3}\exp\left(c_B\sum_{p=1}^{P}|\theta|_p X_p\right)$ can be bounded by an integrable function of X uniformly in a open neighborhood of θ_0.

(A4) The functions $E_{V,X}I(V \geq v)\exp\left(-c_A - c_B\sum_{p=1}^{P}|\theta|_p X_p\right)$ are bounded away from zero uniformly in v and $\eta = (\theta^T, \xi^T)^T$ in a open neighborhood of η_0.

(A5) The matrix $\mathcal{I}$ is positive definite.

Proof of the asymptotic unbiasedness of the estimating equation under the general missing-at-random assumption

In this section, we provide an outline of the proof the asymptotic unbiasedness of the estimating equations $S_\theta = 0$ and $S_\xi = 0$ under the general missing at random mechanism specified by equation (6). Further details of the proof can be found in Supplementary appendix.

First, it is easy to see that the asymptotic limit of $(1/n)S_\theta^{(r)}$ can be written in general form asth

$$E_{R,V,\Delta,Y_o^r}\left(I(R=r)\Delta\left[\frac{\partial}{\partial\theta}\log\{h_{Y^{o_r}}(V|X)\} - \frac{s^{(1)}(V,Y^{o_r})}{s^{(0)}(V,Y^{o_r})}\right]\right),$$

where

$$s^{(1)}(V, Y^{o_r}) = E_{V',X'} I(V' \geq V) \left[\partial \log\{h_{Y^{o_r}}(V'|X')\}/\partial\theta\right] h_{Y^{o_r}}(V'|X'),$$

$s^{(0)}(V, Y^{o_r}) = E_{V',X'} I(V' \geq V) h_{Y^{o_r}}(V'|X')$, and $\partial \log\{h_{Y^{o_r}}(V|X)\}/\partial\theta = \mathcal{E}_{Y^{o_r}}(\mathcal{B}_y|V, X) \otimes X$.

Now, under the missing-at-random assumption, we can show

$$\begin{aligned} C^{(r)} &\equiv E\left[\Delta I(R = r)\partial \log\{h_{Y^{o_r}}(V|X)\}/\partial\theta\right] \\ &= \int_{v,y^{o_r}} E_{V,X}\left(I(V \geq v)\pi^{(r)}(v, X)\left[\partial \log\{h_{y^{o_r}}(v|X)\}/\partial\theta\right] h(v, y^{o_r}|X)\right) dv d\mu(y^{o_r}). \end{aligned}$$

Similarly, we can show,

$$\begin{aligned} D^{(r)} &\equiv E\left\{\Delta I(R = r)\frac{s^{(1)}(V, Y^{o_r})}{s^{(0)}(V, Y^{o_r})}\right\} \\ &= D^{(r)} = \int \frac{s^{(1)}(v, y^{o_r})}{s^{(0)}(v, y^{o_r})} E_{V,X}\left\{I(V \geq v)\pi^{r}(v, X)h_{y^{o_r}}(v|X)\right\} dv d\mu(y^{o_r}). \end{aligned}$$

Now, we note that, if $\pi^{(r)}(T, X) \equiv \pi^{(r)}(T)$, then we have

$$C^{(r)} = D^{(r)} = \int \pi^{(r)}(v) s^{(1)}(v, y^{o_r}) dv d\mu(y^{o_r}),$$

implying the asymptotic unbiasedness of $S_\theta^{(r)}$ for each specific r.

When $\pi^{(r)}(T, X)$ depends on X, in general $C^{(r)} \neq D^{(r)}$, but we can show that

$$\sum_r C^{(r)} = \sum^r D^{(r)} = \int_{v,y} s^{(1)}(v, y) dv d\mu(y)$$

by rearrangements of integrals and sums.

Derivation of the influence function and asymptotic normality

The asymptotic unbiasedness of the estimating function, together with the fact that under the given regularity conditions, $(1/n)\partial T_n/\partial\eta \to \mathcal{I}$ uniformly in an open neighborhood η_0, proves the local consistency of the proposed estimator. In the following lemma, we state a key step that is needed for the proof of Theorem 1.

LEMMA 2.

(1.a) $$n^{1/2}\left\{\frac{S^{(1)}(v,y)}{S^{(0)}(v,y)} - \frac{s^{(1)}(v,y)}{s^{(0)}(v,y)}\right\} =$$

$$\frac{1}{n^{1/2}}\frac{1}{s^{(0)}(v,y)}\sum_{j=1}^n I(V_j \geq v) h_y(V_j|X_j)\left[\frac{\partial}{\partial\theta}\log\{h(V_j, y|X_j)\} - \frac{s^{(1)}(v,y)}{s^{(0)}(v,y)}\right] + o_p(1).$$

(1.b) *The above result holds uniformly for all* (v, y).

The proof of part (a) follows by a standard application of the functional δ-theorem (Theorem 20.8 of van Der Vaart, 1998) by viewing the quantities $S^{(k)}(v,y)$, $k=0,1$, as functions of the underlying empirical process defined by $\{V_j, X_j\}_{j=1}^n$. Since both $s^{(1)}(v,y)$ and $s^{(0)}(v,y)$ are linear functionals of the underlying distribution function of V and X, the required condition for their Hadamard differentiability follows easily under the regularity conditions (A2)-(A3). Moreover, under (A4), we can apply the chain rule to show that the ratio $s^{(1)}(v,y)/s^{(0)}(v,y)$ is Hadamard differentiable. Part (b) of Lemma 2, that is uniform convergence, follows by the uniform boundedness conditions (A2)-(A4).

Now, to derive an asymptotic representation of $n^{-1/2}S_\theta$, we write

$$\begin{aligned}\frac{1}{n^{1/2}}S_\theta &= \frac{1}{n^{1/2}}\sum_r\sum_i \Delta_i I(R_i=r)\left[\frac{\partial}{\partial\theta}\log\{h_{Y_i^{o_r}}(V_i|X_i)\} - \frac{s^{(1)}(V_i,Y_i^{o_r})}{s^{(0)}(V_i,Y_i^{o_r})}\right] \\ &\quad -\frac{1}{n^{1/2}}\sum_r\sum_i \Delta_i I(R_i=r)\left\{\frac{S^{(1)}(V_i,Y_i^{o_r})}{S^{(0)}(V_i,Y_i^{o_r})} - \frac{s^{(1)}(V_i,Y_i^{o_r})}{s^{(0)}(V_i,Y_i^{o_r})}\right\},\end{aligned}$$

where the second term, using Lemma 2, can be written as

$$\sum_r\sum_i \frac{1}{n^{3/2}}\frac{\Delta_i I(R_i=r)}{s^{(0)}(V_i,Y_i^{o_r})}\sum_{j=1}^n I(V_j\geq V_i)h_{Y_i^{o_r}}(V_j|X_j)\left[\frac{\partial}{\partial\theta}\log\{h_{Y_i^{o_r}}(V_j|X_j)\} - \frac{s^{(1)}(V_i,Y_i^{o_r})}{s^{(0)}(V_i,Y_i^{o_r})}\right] + o_p(1),$$

which, in turn, after rearrangement of the sums, can be written as

$$\frac{1}{n^{1/2}}\sum_{j=1}^n\sum_r\left(\frac{1}{n}\sum_{i=1}^n \frac{\Delta_i I(R_i=r)}{s^{(0)}(V_i,Y_i^{o_r})} I(V_i\leq V_j)h_{Y_i^{o_r}}(V_j|X_j)\left[\frac{\partial}{\partial\theta}\log\{h_{Y_i^{o_r}}(V_j|X_j)\} - \frac{s^{(1)}(V_i,Y_i^{o_r})}{s^{(0)}(V_i,Y_i^{o_r})}\right]\right) + o_p(1).$$

Now, by the law of large numbers, it is easy to see that the expression within the above square brackets converges to the second term in the expression of J_{ni}^θ given in Theorem 2. The derivation of the asymptotic representation of S_ξ as an i.i.d. sum requires a similar step. The asymptotic normality of $(\hat\theta_n, \hat\xi_n)$ now follows by the application of the central limit theorem.

Table 1: *Results of the simulation study, where the disease has four subtypes based on two disease traits, each with two levels. The true value of* $\theta^{(0)} = 0.35$, $\theta^{(1)}_{1(2)} = 0$ *and* $\theta^{(1)}_{2(2)} = 0.25$. *Each of the disease traits is missing via a missing-completely-at-random and missing-at-random mechanism. The cohort size for the simulation is* $n = 10,000$. *Evar and 95%CP stand for means of estimated variances and 95% coverage probability, respectively, over the different simulations.*

Method		$\theta^{(0)}$	$\theta^{(1)}_{1(2)}$	$\theta^{(1)}_{2(2)}$	$\theta^{(0)}$	$\theta^{(1)}_{1(2)}$	$\theta^{(1)}_{2(2)}$
		missing-completely-at-random			missing-at-random		
Full-cohort	Bias($\times 10^2$)	-0.06	-0.10	0.42	-0.06	-0.10	0.42
	Var($\times 10^2$)	0.25	0.45	1.15	0.25	0.44	1.15
	Etvar($\times 10^2$)	0.24	0.46	1.15	0.24	0.46	1.15
	95% CP	95.6	94.4	94.2	95.6	94.4	94.2
		20% missing			20% missing		
Complete-	Bias($\times 10^2$)	0.69	-0.13	0.93	2.49	-0.19	0.47
case	Var($\times 10^2$)	0.36	0.65	1.79	0.48	0.98	2.39
	Etvar($\times 10^2$)	0.38	0.72	1.81	0.51	0.97	2.41
	95% CP	96.2	96.4	95.8	94.0	94.6	93.8
Estimating-	Bias($\times 10^2$)	-0.08	-0.11	0.70	0.00	-0.18	0.17
equation	Var($\times 10^2$)	0.26	0.51	1.35	0.28	0.59	1.54
	Etvar($\times 10^2$)	0.27	0.57	1.40	0.29	0.65	1.61
	95% CP	95.6	96.0	96.0	95.8	96.0	96.2
		50% missing			50% missing		
Complete-	Bias($\times 10^2$)	1.24	-0.27	0.03	24.30	1.00	2.90
case	Var($\times 10^2$)	1.06	1.96	5.03	4.48	9.96	22.52
	Etvar($\times 10^2$)	0.98	1.87	4.81	4.29	8.25	24.17
	95% CP	93.4	94.2	94.6	76.8	92.2	95.2
Estimating-	Bias($\times 10^2$)	-0.10	-0.15	0.65	0.10	0.51	-2.66
equation	Var($\times 10^2$)	0.43	1.02	2.57	0.67	1.84	4.55
	Etvar($\times 10^2$)	0.37	0.90	2.18	0.63	1.86	4.47
	95% CP	93.2	93.2	92.0	94.4	93.8	91.6

Table 2: *Results of the simulation study with a misspecified model for the baseline hazard functions. Here, the disease has four subtypes based on two disease traits, each with two levels. The true value of* $\theta^{(0)} = 0.35$, $\theta^{(1)}_{1(2)} = 0$, *and* $\theta^{(1)}_{2(2)} = 0.25$. *Each of the disease traits is missing-completely-at-random and missing-at-random with probability 0.5 or 0.2. The cohort size for the simulation is* $n = 10,000$. *Etvar and 95%CP stand for means of estimated variances and 95% coverage probability, respectively, over the different simulations.*

Method		missing-completely-at-random			missing-at-random		
		$\theta^{(0)}$	$\theta^{(1)}_{1(2)}$	$\theta^{(1)}_{2(2)}$	$\theta^{(0)}$	$\theta^{(1)}_{1(2)}$	$\theta^{(1)}_{2(2)}$
Full-cohort	Bias($\times 10^2$)	0.09	-0.15	0.41	-0.04	0.43	-0.13
	Var($\times 10^2$)	0.14	0.56	0.70	0.15	0.58	0.75
	Etvar($\times 10^2$)	0.14	0.55	0.72	0.14	0.55	0.72
	95% CP	94.8	96.2	95.6	96.0	95.4	94.8
		20% missing			20% missing		
Complete-	Bias($\times 10^2$)	1.15	0.01	0.34	3.11	1.07	-1.37
case	Var($\times 10^2$)	0.20	0.88	1.09	0.30	1.17	1.46
	Etvar($\times 10^2$)	0.22	0.87	1.13	0.29	1.17	1.53
	95% CP	95.4	94.4	94.2	90.8	95.4	95.8
Estimating-	Bias($\times 10^2$)	-0.37	2.02	-0.35	-0.68	3.60	-1.46
equation	Var($\times 10^2$)	0.16	0.65	0.88	0.17	0.74	0.95
	Etvar($\times 10^2$)	0.15	0.64	0.83	0.16	0.72	0.93
	95% CP	93.6	95.0	94.2	94.2	91.4	94.2
		50% missing			50% missing		
Complete-	Bias($\times 10^2$)	2.21	-0.45	0.92	25.84	3.75	-8.37
case	Var($\times 10^2$)	0.59	2.32	2.68	2.66	9.73	14.91
	Etvar($\times 10^2$)	0.56	2.25	2.94	2.35	9.70	13.67
	95% CP	93.6	95.2	96.0	60.6	95.4	92.8
Estimating-	Bias($\times 10^2$)	-0.14	4.49	-0.79	-1.99	7.23	-4.15
equation	Var($\times 10^2$)	0.19	0.97	1.16	1.59	8.82	3.09
	Etvar($\times 10^2$)	0.20	0.96	1.23	1.66	7.23	2.46
	95% CP	94.2	93.4	95.2	91.9	88.2	93.3

Table 3: *Results of the data analysis. Here we consider five disease traits: Grade, Stage, Histology, ER status, and PR status. Grade and Histology each have 3 levels, and Stage, ER status, and PR status each have 2 levels. Here EST and SE stand for estimate and standard error, respectively.*

			Grade (Ref: Well)		Stage (Ref: Localized)	Histology (Ref: Ductal)		ER Status (Ref: ER+)	PR Status (Ref: PR+)
			Moderate	Poor	Distant	Lobular	Other	ER-	PR-
		% missing	24.5		2.0	0.0		32.8	36.6
Method		$\theta^{(0)}$	$\theta^{(1)}_{1(2)}$	$\theta^{(1)}_{1(3)}$	$\theta^{(1)}_{2(2)}$	$\theta^{(1)}_{3(2)}$	$\theta^{(1)}_{3(3)}$	$\theta^{(1)}_{4(2)}$	$\theta^{(1)}_{5(2)}$
Complete-	EST	1.0821	-0.2104	-0.0591	0.8092	-0.2568	0.3762	0.3747	-0.9210
case	SE	0.3051	0.3606	0.4031	0.3120	0.4137	0.4724	0.5062	0.3978
	p-value	0.0004	0.5596	0.8834	0.0095	0.5348	0.4258	0.4591	0.0206
Estimating-	EST	0.8768	0.0749	0.1680	0.6931	-0.5196	0.3269	0.7601	-1.1477
equation	SE	0.2463	0.2893	0.3005	0.2096	0.2741	0.2867	0.4077	0.3409
	p-value	0.0004	0.7957	0.5761	0.0009	0.0580	0.2542	0.0623	0.0008

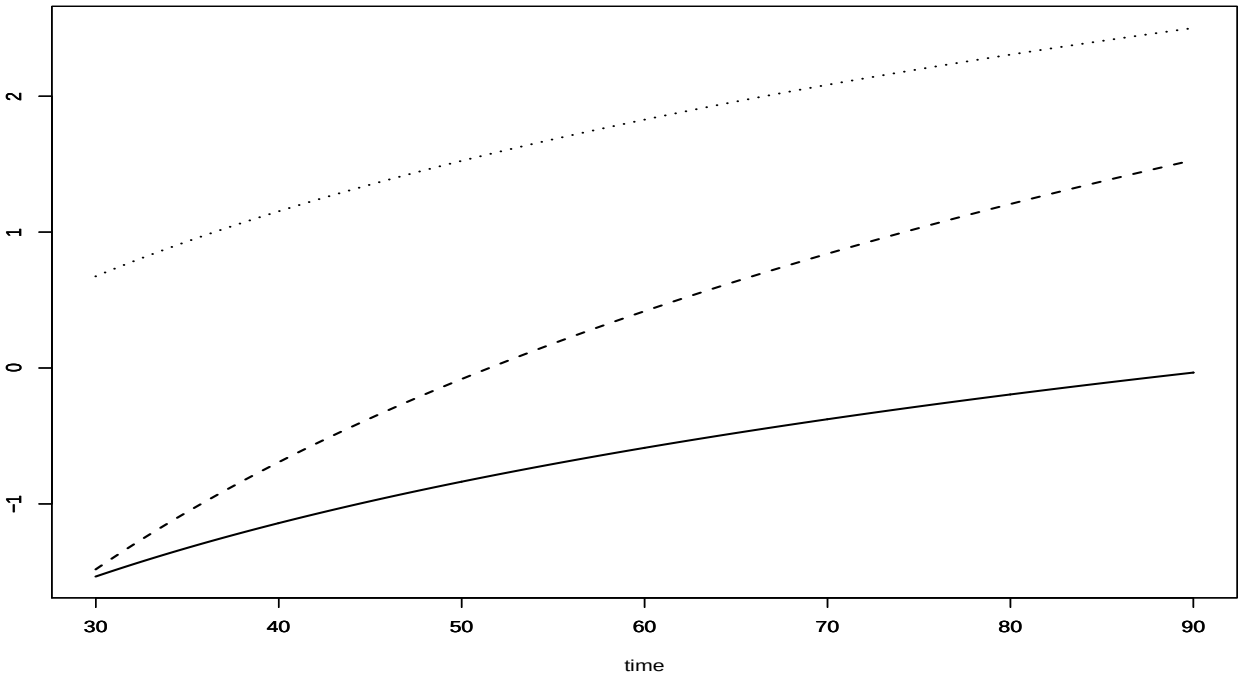

Figure 1: Plot of the $\log\{\lambda_{(y_1,y_2)}(t)/\lambda_{(1,1)}(t)\}$. The solid line (—), dashed line (−−), and dotted line $(\cdots)$ correspond to $(y_1, y_2) = (1, 2)$, $(2, 1)$ and $(2, 2)$, respectively.