Analysis of cohort studies with multivariate and partially observed disease classification data

By NILANJAN CHATTERJEE

Division of Cancer Epidemiology and Genetics, National Cancer Institute,
National Institute of Health. Rockville, Maryland 20852, U.S.A.
chattern@mail.nih.gov

SAMIRAN SINHA

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
sinha@stat.tamu.edu

W. RYAN DIVER AND HEATHER SPENCER FEIGELSON

Department of Epidemiology and Surveillance Research, American Cancer Society,
Atlanta, Georgia 30303, U.S.A.

SUMMARY

Complex diseases like cancers 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 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 cancers. 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 aim of studying etiological 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 difficulties. First, a combination of multiple disease traits, some of which can be ordinal or even continuous, may 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. 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 using 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 hazards 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 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 simulation studies involving small and large numbers of disease subtypes. Moreover, we apply the proposed method to a dataset on breast cancer incidence and various histopathologic traits of breast tumours from the well-known Cancer Prevention Study II of the American Cancer Society.

2. Model and assumption

2·1. *Notation and models for the cause-specific hazard*

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. 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 tumour characteristics. If the kth trait defines M_k categories for the disease, then 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. Suppose that $\mathcal{D}_Y$ denotes the set of all possible values of Y. Breast cancer patients, for example, are often classified into four categories based on the presence or absence of estrogen and progesterone receptors. Given that follow-up ends at the occurrence of

any type of disease, the M subtypes can be treated as competing events. If X denotes a 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

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

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 can become very large. In the following section, we consider reducing the number of these parameters by using a second-stage model, following an idea introduced by Chatterjee (2004) in the context of polytomous logistic regressions.

2·2. *A second-stage regression model for the subtype-specific regression parameter*

First, we focus on modelling the regression coefficients associated with a single covariate. 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:

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

where $\theta^{(0)}$ represents the regression coefficient corresponding to a referent subtype of the disease, the coefficients $\theta_{k(y_k)}^{(1)}$ represent the first-order parameter contrasts, $\theta_{kk'(y_k,y_{k'})}^{(2)}$ 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

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

and in this case the heterogeneity between two subtype-specific regression coefficients is $\beta_{(y_1,\ldots,y_{k-1},y_k,y_{k+1},\ldots,y_K)} - \beta_{(y_1,\ldots,y_{k-1},y_k^*,y_{k+1},\ldots,y_K)} = \theta_{k(y_k)}^{(1)} - \theta_{k(y_k^*)}^{(1)}$. Thus, in this model, the contrasts of the form $\theta_{k(y_k)}^{(1)} - \theta_{k(y_k^*)}^{(1)}$ give a measure of the degree of etiologic heterogeneity among the disease subtypes with respect to the kth 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

$$\theta_{k(y_k)}^{(1)} = \theta_k^{(1)} s_{y_k}^{(k)} \quad (k = 1, \ldots, M_k),$$

where $\{0 = s_1^{(k)} \leqslant s_2^{(k)} \cdots \leqslant s_{M_k}^{(k)}\}$ are a set of prespecified scores assigned to the M_k levels of the kth trait. This model summarizes the degree of etiologic heterogeneity with respect to the kth 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 to include interaction terms between pairs of disease traits, thus potentially allowing the degree of etiologic heterogeneity for 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 modelling them in terms of underlying continuous scores. Details of modelling may 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 β 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 for notational convenience that $\mathcal{B}^{(p)} = \mathcal{B}$ for all p.

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

$$L_n = \prod_{i:\Delta_i=1} \left\{ \frac{h_{Y_i}(V_i \mid X_i)}{\sum_{j=1}^n I(V_j \geqslant V_i) h_{Y_i}(V_i \mid X_j)} \right\}$$

$$= \prod_{i:\Delta_i=1} \left\{ \frac{\exp\left(\sum_{p=1}^P \mathcal{B}_{Y_i}^{\mathrm{T}} \theta_p X_{ip} \right)}{\sum_{j=1}^n I(V_j \geqslant V_i) \exp\left(\sum_{p=1}^P \mathcal{B}_{Y_i}^{\mathrm{T}} \theta_p X_{jp} \right)} \right\}, \tag{3}$$

where $Y_i = (Y_{i1}, \ldots, Y_{iK})$ denotes the observed disease traits for the ith subject given $\Delta_i = 1$, i.e., for a case, and $\mathcal{B}_{Y_i}^{\mathrm{T}}$ 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 (3) without requiring any assumption about the subtype-specific baseline hazard functions $\lambda_y(t)$. The score function associated with (3) can be written as

$$S_\theta^* = \sum_{\Delta_i=1} \left\{ \mathcal{B}_{Y_i} \otimes X_i - \frac{\sum_{j=1}^n I(V_j \geqslant V_i) \mathcal{B}_{Y_i} \otimes X_j \exp\left(\sum_{p=1}^P \mathcal{B}_{Y_i}^{\mathrm{T}} \theta_p X_{jp} \right)}{\sum_{j=1}^n I(V_j \geqslant V_i) \exp\left(\sum_{p=1}^P \mathcal{B}_{Y_i}^{\mathrm{T}} \theta_p X_{jp} \right)} \right\},$$

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 (3) to perform a complete-case analysis involving only those cases that have complete data on all the disease traits. Such complete-case analyses, however, may potentially require discarding large numbers of cases, resulting in inevitable loss of efficiency and possibly bias due to nonrandom missingness.

Several researchers 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) 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 modelling of the baseline hazard functions. Lu & Tsiatis (2005) 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, that of Goetghebeur & Ryan (1995) is particularly apt for extension to our setting, as it 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 simplifies to the standard partial likelihood method. When there are missing data, though

some assumptions about the inter-relationship 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 notation. Let R_k denote the indicator of whether the kth 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 $r = (r_1, \ldots, r_K)$ denote 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. Suppose that $\mathcal{D}_R$ denotes the set of all possible values of R. We will assume that the data are missing-at-random, in the sense that

$$\mathrm{pr}(R = r \mid T, X, Y, \Delta = 1) = \mathrm{pr}(R = r \mid T, X, \Delta = 1) = \pi^{(r)}(T, X), \tag{4}$$

i.e. the probabilities for the different patterns of missing data do not depend on the trait values Y. The model (4) 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

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

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

Define

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

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

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

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

$$S_\theta \equiv \sum_{r \in \mathcal{D}_R} S_\theta^{(r)} \equiv \sum_{r \in \mathcal{D}_R} \sum_{\Delta_i = 1, R_i = r} \left\{ \mathcal{E}_{Y_i^{o_r}}(\mathcal{B}_Y \mid T_i, X_i) \otimes X_i \right.$$
$$\left. - \frac{\sum_{j=1}^n I(V_j \geqslant V_i) \mathcal{E}_{Y_i^{o_r}}(\mathcal{B}_Y \mid T_i, X_j) \otimes X_j h_{Y_i^{o_r}}(T_i \mid X_j)}{\sum_{j=1}^n I(V_j \geqslant V_i) h_{Y_i^{o_r}}(T_i \mid X_j).} \right\} = 0. \tag{5}$$

The unbiasedness of the left-hand side of (5) under the general missing-at-random model (4) is proven in the Appendix. In the absence of any missing data on disease traits, $\mathcal{E}_{Y_i^{o_r}}(\mathcal{B}_Y \mid X)$ corresponds to the row of the second-stage design matrix $\mathcal{B}$ associated with the observed disease traits Y_i and (5) is equivalent to the score equation associated with the partial likelihood (3). If we assumed discrete disease subtypes and did not impose any structure through the second-stage model, then $\mathcal{E}_{y_i^o}(\mathcal{B}_Y \mid X_i, T_i)$ would simply correspond to the conditional probability vector for

observing the different disease subtypes, given the observed traits, and (5) would essentially be the equation proposed by Goetghebeur & Ryan for dealing with two failure types.

The estimating equation (5) 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 becomes large, the number of cases for the individual subtypes becomes sparse. In the following, we consider a semiparametric estimation approach for the baseline hazard functions. First, we express

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

where $\lambda_{(1,\dots,1)}(t)$ 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 & 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

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

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

$$h_u(t \mid X) = \int_{y \in \mathcal{D}_Y} \lambda_y(t_i)\exp(X_i\beta_y)d\mu(y),$$

$$Q_u^y(X) = \frac{\exp(\alpha_y + X\beta_y)}{\int_{y \in \mathcal{D}_Y}\exp(\alpha_y + X\beta_y)d\mu(y)}, \quad w_u(X) = \int_{y \in \mathcal{D}_Y}\exp(\alpha_y + X\beta_y)d\mu(y)$$

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

Under the constant-hazard-ratio assumption $\alpha_y(t) \equiv \alpha_y$, the quantities $Q_{y^{or}}^{y^{m_r}}(t, X)$ and $h_{y^{or}}(t \mid X)$ in the definition of (5) can be replaced by

$$Q_{y^{or}}^{y^{m_r}}(X) = \frac{\exp\big(\alpha_{(y^{m_r}, y^{or})} + X\beta_{(y^{m_r}, y^{or})}\big)}{\int_{y^{m_r}}\exp\big(\alpha_{(y^{or}, y^{m_r})} + X\beta_{(y^{or}, y^{m_r})}\big)d\mu(y^{m_r})}$$

and

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

respectively. We further let

$$\mathcal{E}_{y^{or}}(\mathcal{A}_Y \mid X) = \int_{y^{m_r}}\mathcal{A}_{(y^{or}, y^{m_r})}Q_{y^{or}}^{y^{m_r}}(X)d\mu(y^{m_r}) \text{ and } \mathcal{E}_u(\mathcal{A}_Y \mid X) = \int_{y \in \mathcal{D}_Y}\mathcal{A}_y Q_u^y(X)d\mu(y).$$

Now, following Goetghebeur and Ryan (1995), we propose estimating ξ based on the partial likelihood

$$L_n^* = \prod_{r \in \mathcal{D}_R}\prod_{R_i=r, \Delta_i=1}\left\{\frac{h_{y_i^{or}}(V_i \mid X_i)}{\sum_{j=1}^{n} I(V_j \geqslant V_i)h_u(V_i \mid X_j)}\right\}.$$

The associated score function can be conveniently expressed as

$$S_\xi \equiv \sum_{r \in \mathcal{D}_R} S_\xi^{(r)} \equiv \sum_{r \in \mathcal{D}_R} \sum_{\Delta_i = 1, R_i = r} \left\{ \mathcal{E}_{y_i^{o_r}}(\mathcal{A}_Y \mid X_i) - \frac{\sum_{j=1}^n I(V_j \geqslant V_i)\mathcal{E}_u(\mathcal{A}_Y \mid X_j)\omega_u(X_j)}{\sum_{j=1}^n I(V_j \geqslant V_i)\omega_u(X_j)} \right\}.$$

In our applications, we solved both sets of estimating equations $S_\theta = 0$ and $S_\xi = 0$ using the Newton–Raphson method. In principle, these equations can also be solved by an EM-type algorithm where the expectation steps will involve computing the conditional expectations $\mathcal{E}_{y_i^O}(\mathcal{B}_Y \mid X_i, T_i)$ and $\mathcal{E}_{y_i^{o_r}}(\mathcal{A}_Y \mid X_i)$, and then the M-step will involve solving partial-likelihood-type score equations of the form (5).

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 & 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 misspecification 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* (4)*, as $n \to \infty$, the convergence results*

$$n^{-1} S_\theta \equiv n^{-1} \sum_{r \in \mathcal{D}_R} S_\theta^{(r)} \to 0, \quad n^{-1} S_\xi \equiv n^{-1} \sum_{r \in \mathcal{D}_R} S_\xi^{(r)} \to 0$$

hold in probability at the true parameter values. Moreover, if $\pi(T, X) = \pi(T)$, then for each specific r,

$$n^{-1} S_\theta^{(r)} \to 0 \quad and \quad n^{-1} S_\xi^{(r)} \to 0,$$

in probability.

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 invalid under the more general missing data model we consider. Now, we define some further notation. Let

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

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

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

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 $\eta = (\theta^{\mathrm{T}}, \xi^{\mathrm{T}})^{\mathrm{T}}$ is estimated by solving

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

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*

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

where

$$J_{ni}^\theta = \Delta_i \sum_{r\in\mathcal{D}_R} I(R_i = r)\left\{ \mathcal{E}_{Y_i^{o_r}}(\mathcal{B}_Y \mid X_i) \otimes X_i - \frac{s^{(1)}(V_i, Y_i^{o_r})}{s^{(0)}(V_i, Y_i^{o_r})} \right\}$$

$$+ \sum_{r\in\mathcal{D}_R} E_{\Delta, R, V, Y^{o_r}}\left[\Delta I(R = r)\frac{I(V \leqslant V_i)\omega_{Y^{o_r}}(X_i)}{s^{(0)}(V, Y^{o_r})}\left\{ \mathcal{E}_{Y^{o_r}}(\mathcal{B}_Y \mid X_i) \otimes X_i - \frac{s^{(1)}(V, Y^{o_r})}{s^{(0)}(V, Y^{o_r})} \right\} \right]$$

and

$$J_{ni}^\xi = \Delta_i \sum_{r\in\mathcal{D}_R} I(R_i = r)\left\{ \mathcal{E}_{Y_i^{o_r}}(\mathcal{A}_Y \mid X_i) \otimes X_i - \frac{s^{(1)}(V_i, u)}{s^{(0)}(V_i, u)} \right\}$$

$$+ E_{\Delta, V}\left[\Delta\frac{I(V \leqslant V_i)\omega_u(X_i)}{s^{(0)}(V, u)}\left\{ \mathcal{E}_u(\mathcal{A}_Y \mid X_i) - \frac{s^{(1)}(V, u)}{s^{(0)}(V, u)} \right\} \right].$$

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

$$\hat{J}_{ni}^\theta = \Delta_i \sum_{r\in\mathcal{D}_R} I(R_i = r)\left\{ \mathcal{E}_{Y_i^{o_r}}(\mathcal{B}_Y \mid X_i) \otimes X_i - \frac{S^{(1)}(V_i, Y_i^{o_r})}{S^{(0)}(V_i, Y_i^{o_r})} \right\}$$

$$+ \sum_{r\in\mathcal{D}_R} \sum_{l=1}^n \Delta_l I(R_l = r)\frac{I(V_l \leqslant V_i)\omega_{Y_l^{o_r}}(X_i)}{S^{(0)}(V_l, Y_l^{o_r})}\left\{ \mathcal{E}_{Y_l^{o_r}}(\mathcal{B}_Y \mid X_i) \otimes X_i - \frac{S^{(1)}(V_l, Y_l^{o_r})}{S^{(0)}(V_l, Y_l^{o_r})} \right\}$$

and

$$\hat{J}_{ni}^\xi = \Delta_i \sum_{r\in\mathcal{D}_R} I(R_i = r)\left\{ \mathcal{E}_{Y_i^{o_r}}(\mathcal{A}_Y \mid X_i) \otimes X_i - \frac{S^{(1)}(V_i, u)}{S^{(0)}(V_i, u)} \right\}$$

$$+ \sum_{l=1}^n \Delta_l \frac{I(V_l \leqslant V_i)\omega_u(X_i)}{S^{(0)}(V_l, u)}\left\{ \mathcal{E}_u(\mathcal{A}_Y \mid X_i) - \frac{S^{(1)}(V_l, u)}{S^{(0)}(V_l, u)} \right\}.$$

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

4. SIMULATION STUDY

4·1. *Background information*

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. 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.

4·2. *Moderate number of disease subtypes*

In this setting, we simulated data by mimicking the incidence pattern of four subtypes of breast cancer defined by the presence or absence of estrogen and progesterone receptors. 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 $N(0, \sigma = 1)$ 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

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

with a Weibull specification for the baseline hazard functions,

$$\lambda_{(y_1,y_2)}(t) = \gamma_{(y_1,y_2)}\lambda_{(y_1,y_2)}^{\gamma_{(y_1,y_2)}}t^{\gamma_{(y_1,y_2)}-1} \quad (y_1, y_2 = 1, 2). \tag{6}$$

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, indicating 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 progesterone receptor positive tumours compared with progesterone receptor negative tumours. We assumed $\theta^{(1)}_{1(2)} = 0$, that is, no difference in the effect of the covariate between estrogen-receptor positive and negative tumours. 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 further specified $\lambda_{(y_1,y_2)}$ 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

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

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 $N(75, 5^2)$ distribution. In our simulation setting, we observed, on average, 3·91%, 3·86%, 0·56% and 0·51% 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 estrogen receptor and progesterone receptor 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 missingness-at-random, we assumed the trait values to be completely observed for cases whose X and T values are greater than the 80th percentile values for the respective distributions. Among the remaining cases, we simulated missing traits by deleting estrogen receptor and progesterone receptor 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.

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 $\theta^{(0)} = 0\cdot35$, $\theta^{(1)}_{1(2)} = 0$ and $\theta^{(1)}_{2(2)} = 0\cdot25$. 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 $n = 10\,000$*

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 ($\times10^2$)	$-0\cdot06$	$-0\cdot10$	$0\cdot42$	$-0\cdot06$	$-0\cdot10$	$0\cdot42$
	Var ($\times10^2$)	$0\cdot25$	$0\cdot45$	$1\cdot15$	$0\cdot25$	$0\cdot44$	$1\cdot15$
	Evar ($\times10^2$)	$0\cdot24$	$0\cdot46$	$1\cdot15$	$0\cdot24$	$0\cdot46$	$1\cdot15$
	95% CP	$95\cdot6$	$94\cdot4$	$94\cdot2$	$95\cdot6$	$94\cdot4$	$94\cdot2$
		20% missing			20% missing		
Complete case	Bias ($\times10^2$)	$0\cdot69$	$-0\cdot13$	$0\cdot93$	$2\cdot49$	$-0\cdot19$	$0\cdot47$
	Var ($\times10^2$)	$0\cdot36$	$0\cdot65$	$1\cdot79$	$0\cdot48$	$0\cdot98$	$2\cdot39$
	Evar ($\times10^2$)	$0\cdot38$	$0\cdot72$	$1\cdot81$	$0\cdot51$	$0\cdot97$	$2\cdot41$
	95% CP	$96\cdot2$	$96\cdot4$	$95\cdot8$	$94\cdot0$	$94\cdot6$	$93\cdot8$
Estimating equation	Bias ($\times10^2$)	$-0\cdot08$	$-0\cdot11$	$0\cdot70$	$0\cdot00$	$-0\cdot18$	$0\cdot17$
	Var ($\times10^2$)	$0\cdot26$	$0\cdot51$	$1\cdot35$	$0\cdot28$	$0\cdot59$	$1\cdot54$
	Evar ($\times10^2$)	$0\cdot27$	$0\cdot57$	$1\cdot40$	$0\cdot29$	$0\cdot65$	$1\cdot61$
	95% CP	$95\cdot6$	$96\cdot0$	$96\cdot0$	$95\cdot8$	$96\cdot0$	$96\cdot2$
		50% missing			50% missing		
Complete case	Bias ($\times10^2$)	$1\cdot24$	$-0\cdot27$	$0\cdot03$	$24\cdot30$	$1\cdot00$	$2\cdot90$
	Var ($\times10^2$)	$1\cdot06$	$1\cdot96$	$5\cdot03$	$4\cdot48$	$9\cdot96$	$22\cdot52$
	Evar ($\times10^2$)	$0\cdot98$	$1\cdot87$	$4\cdot81$	$4\cdot29$	$8\cdot25$	$24\cdot17$
	95% CP	$93\cdot4$	$94\cdot2$	$94\cdot6$	$76\cdot8$	$92\cdot2$	$95\cdot2$
Estimating equation	Bias ($\times10^2$)	$-0\cdot10$	$-0\cdot15$	$0\cdot65$	$0\cdot10$	$0\cdot51$	$-2\cdot66$
	Var ($\times10^2$)	$0\cdot43$	$1\cdot02$	$2\cdot57$	$0\cdot67$	$1\cdot84$	$4\cdot55$
	Evar ($\times10^2$)	$0\cdot37$	$0\cdot90$	$2\cdot18$	$0\cdot63$	$1\cdot86$	$4\cdot47$
	95% CP	$93\cdot2$	$93\cdot2$	$92\cdot0$	$94\cdot4$	$93\cdot8$	$91\cdot6$

Evar, means of estimated variances; 95% CP, 95% coverage probability.

Under each scenario, we simulated 500 datasets and analyzed each of them using three different methods. In the first, subsequently referred to as full cohort analysis, we applied the partial likelihood (3) 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 (3) to the cohort after deleting all the cases that 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 traits missing.

Table 1 reveals that the proposed estimating equation method had negligible bias for estimation of both the null and nonnull 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 the missingness-completely-at-random mechanism, 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 data was modest, e.g. 20%, the estimating equation method lost only 4–15% efficiency compared with the full-cohort analysis. In contrast, even with a modest amount of missing data, the loss of efficiency for the complete-case analysis ranged between 30% and 40%. Under the missing-at-random mechanism,

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 $\theta^{(0)} = 0.35$, $\theta_{1(2)}^{(1)} = 0$, and $\theta_{2(2)}^{(1)} = 0.25$. 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 $n = 10\,000$*

Method		Missing-completely-at-random			Missing-at-random		
		$\theta^{(0)}$	$\theta_{1(2)}^{(1)}$	$\theta_{2(2)}^{(1)}$	$\theta^{(0)}$	$\theta_{1(2)}^{(1)}$	$\theta_{2(2)}^{(1)}$
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
	Evar $(\times 10^2)$	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 case	Bias $(\times 10^2)$	1·15	0·01	0·34	3·11	1·07	−1·37
	Var $(\times 10^2)$	0·20	0·88	1·09	0·30	1·17	1·46
	Evar $(\times 10^2)$	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 equation	Bias $(\times 10^2)$	−0·37	2·02	−0·35	−0·68	3·60	−1·46
	Var $(\times 10^2)$	0·16	0·65	0·88	0·17	0·74	0·95
	Evar $(\times 10^2)$	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 case	Bias $(\times 10^2)$	2·21	−0·45	0·92	25·84	3·75	−8·37
	Var $(\times 10^2)$	0·59	2·32	2·68	2·66	9·73	14·91
	Evar $(\times 10^2)$	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 equation	Bias $(\times 10^2)$	−0·14	4·49	−0·79	−1·99	7·23	−4·15
	Var $(\times 10^2)$	0·19	0·97	1·16	1·59	8·82	3·09
	Evar $(\times 10^2)$	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

Evar, means of estimated variances; 95% CP, 95% coverage probability.

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 the missing-at-random mechanism, the loss of efficiency for the complete case analysis was more dramatic than that observed under the missing-completely-at-random mechanism.

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 was the same as above, except that in model (6) 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 $\lambda_{(2,2)} = 0.0072$, resulting in approximately 7.06%, 0.73%, 1.64% and 1.34% of subtypes $(2, 2)$, $(1, 2)$, $(2, 1)$ and $(1, 1)$, respectively, in the underlying cohort. Figure 1 of the Supplementary Material shows how $\log\{\lambda_{(y_1,y_2)}(t)/\lambda_{(1,1)}(t)\}$ changes over time. From the results shown in Table 2, we observe that in this setting, under the missing-completely-at-random mechanism 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 $\theta_{1(2)}^{(1)}$ under the setting of 50% missing data. In contrast, under the missing-at-random mechanism, the complete-case analysis produced severe bias in estimating the parameter $\theta^{(0)}$ and the corresponding 95% coverage probability was unacceptably low. Under the missing-at-random mechanism, the bias of the estimating equation method also

increased, but still remained small in absolute terms, and the corresponding 95% coverage probabilities were reasonable.

4·3. *Large number of disease subtypes*

In the third 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

$$\beta_{(y_1, y_2, y_3)} = \theta^{(0)} + \theta_1^{(1)} s_{1(y_1)} + \theta_2^{(1)} s_{2(y_2)} + \theta_3^{(1)} s_{3(y_3)},$$

with $s_{j(y_j)} = (y_j - 1)^{0·3}$ (Chatterjee, 2004) and the baseline hazard functions followed a Weibull distribution of the form (6). We chose $\theta^{(0)} = 0·35$, $\theta_1^{(1)} = 0·15$, $\theta_2^{(1)} = 0·0$ and $\theta_3^{(1)} = 0·5$. We allowed 64 unrestricted values for the λ- and γ- parameters of the baseline Weibull distribution by randomly drawing their values from the Un(3·5, 4) and Un(0·0021, 0·0024) distributions, respectively. As before, $X \sim N(0, 1)$ and the censoring time was generated from an $N(75, 5^2)$ 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, $n = 5000$ and $10\,000$.

As before, we analyzed each dataset 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 $\log\{\alpha_{(y_1, y_2, y_3)}\} = \xi^{(0)} + \xi_{1(y_1)}^{(1)} + \xi_{2(y_2)}^{(2)} + \xi_{3(y_3)}^{(3)}$. The results from this simulation study, shown in Table 1 of the Supplementary Material, 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. Furthermore, 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 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 30 June 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, behavioural, 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 follow-up questionnaires were at least 90% (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 the risk of breast cancer in the Cancer Prevention Study II cohort. After excluding those women who were 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 at postmenopausal baseline in 1992.

Among the 44 172 women, we found that 1516 had some form of breast cancer. The cancer cases were verified either by obtaining medical records or through linkage with state cancer registries when complete medical records could not be obtained. We analyze available data

Table 3. *Results of the data analysis. Here we consider five disease traits: grade, stage, histology, estrogen receptor status and progesteron receptor status. Grade and histology each have three levels, and stage, estrogen receptor status, and progesteron receptor status each have two levels*

Method		% missing $\theta^{(0)}$	Grade (Ref.: Well) Moderate 24.5 $\theta^{(1)}_{1(2)}$	Poor $\theta^{(1)}_{1(3)}$	Stage (Ref.: Localized) Distant 2.0 $\theta^{(1)}_{2(2)}$	Histology (Ref.: Ductal) Lobular 0.0 $\theta^{(1)}_{3(2)}$	Other $\theta^{(1)}_{3(3)}$	ER Status (Ref.: ER+) ER− 32.8 $\theta^{(1)}_{4(2)}$	PR Status (Ref.: PR+) PR− 36.6 $\theta^{(1)}_{5(2)}$
Complete	EST	1·082	−0·210	−0·059	0.809	−0·257	0.376	0.375	−0·921
case	SE	0.305	0·361	0·403	0.312	0·414	0·472	0.506	0·398
	p-value	<0·001	0·559	0·883	0.009	0·535	0.426	0.459	0·021
Estimating	EST	0.877	0.075	0·168	0.693	−0·519	0.327	0.760	−1·148
equation	SE	0.246	0.289	0·301	0.209	0·274	0·287	0.408	0·341
	p-value	<0·001	0·796	0·576	0.001	0·058	0.254	0.062	0·001

EST, estimate; SE, standard error; ER, estrogen receptor; PR, progesteron receptor.

on five tumour traits: grade, with three categories, well/moderately/poorly differentiated; stage, with two categories, localized/distant; histologic type, with three categories, ductal/lobular/other; estrogen receptor status with two categories, positive/negative; progesterone receptor status, with two categories, positive/negative. The aim of our analysis was to study how the association between weight gain and risk of breast cancer varied by various tumour 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 had information on all of the disease traits, while the remaining 734 subjects had information missing on at least one of the traits. Let y_1, y_2, y_3, y_4 and y_5 denote the level of grade, stage, histology, estrogen receptor status, and progesterone receptor status, respectively. We modelled the hazard of the various cancer subtypes as

$$h_{(y_1,y_2,y_3,y_4,y_5)}(t \mid X) = \lambda_{(y_1,y_2,y_3,y_4,y_5)}(t) \exp\big(X\beta_{(y_1,y_2,y_3,y_4,y_5)}\big),$$

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, estrogen receptor positive and progesterone receptor positive as the referent levels for the associated tumour 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

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

From the results in Table 3, it is evident that both methods produced estimates of $\theta^{(0)}$ that are 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 tumours. The significance of the estimate of $\theta^{(1)}_{5(2)}$ in both methods also suggested that the association between weight gain and risk of breast cancer is stronger for progesterone receptor positive compared with progesterone receptor negative tumours. For all parameters, the estimating equation method produces substantially smaller standard errors than does the complete-case analysis.

6. Discussion

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 modelling 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 inter-relationships of the nuisance baseline hazard functions for the different disease subtypes.

Our simulation study suggests that the bias generated by the proposed method for estimation of θ, the focus parameters of interest, is generally quite small, even when the model for $\lambda_y(t)/\lambda_{(1,...,1)}(t)$ is grossly misspecified. Our theory is valid for more general parametric specification of $\lambda_y(t)/\lambda_{(1,...,1)}(t)$ 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 use the constant hazard-ratio and log-linear modelling assumption to construct a doubly robust estimator for θ (Gao & Tsiatis, 2005) that would be unbiased for a correctly specified model either for $\pi(T, X)$ or for $\lambda_y(t)/\lambda_{(1,...,1)}(t)$. 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. The computer code is available from the authors upon request.

Supplementary material

Supplementary material is available at *Biometrika* online.

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.

(C1) The missingness probability $\pi^{(r)}(T, S) > 0$ almost surely in (T, S) for $r = (1, \ldots, 1)$.
(C2) 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.
(C3) Assume that the function $X^{\otimes 3} \exp(c_B \sum_{p=1}^P |\theta|_p X_p)$ can be bounded by an integrable function of X uniformly in an open neighbourhood of θ_0.
(C4) The functions $E_{V,X} I(V \geqslant v) \exp(-c_A - c_B \sum_{p=1}^P |\theta|_p X_p)$ are bounded away from zero uniformly in v and $\eta = (\theta^\mathsf{T}, \xi^\mathsf{T})^\mathsf{T}$ in an open neighbourhood of η_0.
(C5) The matrix $\mathcal{I}$ is positive definite.

Proof of the asymptotic unbiasedness of the estimating equation under the general missing-at-random assumption. 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 (4). Further details of the proof can be found in the Supplementary Material.

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

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

where

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

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

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

$$C^{(r)} \equiv E\left[\Delta I(R = r)\partial \log\{h_{Y^{o_r}}(V \mid X)\}/\partial\theta\right]$$

$$= \int_{v,y^{o_r}} E_{V,X}\left(I(V \geqslant v)\pi^{(r)}(v, X)[\partial \log\{h_{y^{o_r}}(v \mid X)\}/\partial\theta]h(v, y^{o_r} \mid X)\right)dvd\mu(y^{o_r}).$$

Similarly, we can show,

$$D^{(r)} \equiv E\left\{\Delta I(R = r)\frac{s^{(1)}(V, Y^{o_r})}{s^{(0)}(V, Y^{o_r})}\right\}$$

$$= D^{(r)} = \int \frac{s^{(1)}(v, y^{o_r})}{s^{(0)}(v, y^{o_r})} E_{V,X}\{I(V \geqslant v)\pi^r(v, X)h_{y^{o_r}}(v \mid X)\}dvd\mu(y^{o_r}).$$

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

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

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

$$\sum_{r \in \mathcal{D}_R} C^{(r)} = \sum_{r \in \mathcal{D}_R} D^{(r)} = \int_{v,y} s^{(1)}(v, y)dvd\mu(y)$$

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 neighbourhood 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 A1. *Under the regularity conditions described in* (C1)–(C5),

(i) $n^{1/2}\left\{\dfrac{S^{(1)}(v, y)}{S^{(0)}(v, y)} - \dfrac{s^{(1)}(v, y)}{s^{(0)}(v, y)}\right\}$

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

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

The proof of part (i) 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 (C2)–(C3). Moreover, under (C4), we can apply the chain rule to show that the ratio $s^{(1)}(v, y)/s^{(0)}(v, y)$ is Hadamard differentiable. Part (ii) of Lemma A1, that is, uniform convergence, follows by the uniform boundedness conditions (C2)–(C4).

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

$$\frac{1}{n^{1/2}}S_\theta = \frac{1}{n^{1/2}}\sum_{r\in\mathcal{D}_R}\sum_{i=1}^n \Delta_i I(R_i=r)\left[\frac{\partial}{\partial\theta}\log\left\{h_{Y_i^{o_r}}(V_i\mid X_i)\right\} - \frac{s^{(1)}(V_i,Y_i^{o_r})}{s^{(0)}(V_i,Y_i^{o_r})}\right]$$

$$- \frac{1}{n^{1/2}}\sum_{r\in\mathcal{D}_R}\sum_{i=1}^n \Delta_i I(R_i=r)\left\{\frac{S^{(1)}(V_i,Y_i^{o_r})}{S^{(0)}(V_i,Y_i^{o_r})} - \frac{s^{(1)}(V_i,Y_i^{o_r})}{s^{(0)}(V_i,Y_i^{o_r})}\right\},$$

where the second term, using Lemma A1, can be written as

$$\sum_{r\in\mathcal{D}_R}\sum_{i=1}^n \frac{1}{n^{3/2}}\frac{\Delta_i I(R_i=r)}{s^{(0)}(V_i,Y_i^{o_r})}\sum_{j=1}^n I(V_j\geqslant V_i)h_{Y_i^{o_r}}(V_j\mid X_j)\left[\frac{\partial}{\partial\theta}\log\left\{h_{Y_i^{o_r}}(V_j\mid X_j)\right\} - \frac{s^{(1)}(V_i,Y_i^{o_r})}{s^{(0)}(V_i,Y_i^{o_r})}\right] + o_p(1),$$

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

$$\frac{1}{n^{1/2}}\sum_{j=1}^n\sum_{r\in\mathcal{D}_R}\left(\frac{1}{n}\sum_{i=1}^n \frac{\Delta_i I(R_i=r)}{s^{(0)}(V_i,Y_i^{o_r})}I(V_i\leqslant V_j)h_{Y_i^{o_r}}(V_j\mid X_j)\right.$$

$$\left.\times\left[\frac{\partial}{\partial\theta}\log\left\{h_{Y_i^{o_r}}(V_j\mid X_j)\right\} - \frac{s^{(1)}(V_i,Y_i^{o_r})}{s^{(0)}(V_i,Y_i^{o_r})}\right]\right) + o_p(1). \tag{A1}$$

Now, by the law of large numbers, it is easy to see that the expression within the square brackets of expression (A1) 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 a sum of independently and identically distributed random variables requires a similar step. The asymptotic normality of $(\hat\theta_n, \hat\xi_n)$ now follows by the application of the central limit theorem.

References

CHATTERJEE, N. (2004). A two-stage regression model for epidemiological studies with multivariate disease classification data. *J. Am. Statist. Assoc.* **99**, 127–38.

Cox, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc.* 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**, 543–58.

DEWANJI, A. (1992). A note on a test for competing risks with missing failure type. *Biometrika* **79**, 855–7.

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**, 875–91.

GOETGHEBEUR, E. & RYAN, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika* **82**, 821–33.

LU, K. & TSIATIS, A. A. (2005). Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. *Lifetime Data Anal.* **11**, 29–40.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*, Cambridge, UK: Cambridge University Press.