Published in final edited form as:

Lifetime Data Anal. 2018 July ; 24(3): 425–442. doi:10.1007/s10985-017-9401-8.

The competing risks Cox model with auxiliary case covariates under weaker missing-at-random cause of failure

Daniel Nevo,

Departments of Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Reiko Nishihara,

Departments of Epidemiology and Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Shuji Ogino, and

Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

Division of MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Womens Hospital and Harvard Medical School, Boston, MA, USA

Molin Wang

Departments of Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Abstract

In the analysis of time-to-event data with multiple causes using a competing risks Cox model, often the cause of failure is unknown for some of the cases. The probability of a missing cause is typically assumed to be independent of the cause given the time of the event and covariates measured before the event occurred. In practice, however, the underlying missing-at-random assumption does not necessarily hold. Motivated by colorectal cancer molecular pathological epidemiology analysis, we develop a method to conduct valid analysis when additional auxiliary variables are available for cases only. We consider a weaker missing-at-random assumption, with missing pattern depending on the observed quantities, which include the auxiliary covariates. We use an informative likelihood approach that will yield consistent estimates even when the underlying model for missing cause of failure is misspecified. The superiority of our method over naive methods in finite samples is demonstrated by simulation study results. We illustrate the use of our method in an analysis of colorectal cancer data from the Nurses' Health Study cohort, where, apparently, the traditional missing-at-random assumption fails to hold.

Keywords

Competing risks; Masked cause of failure; Missing-at-Random; Subtype analysis

1 Introduction

Disease heterogeneity is commonly represented by molecular subtyping of the disease. When considering time-to-disease data, each of the subtypes can be considered as a cause for the event of interest, the disease. Those subtypes, typically defined by biomarkers, are used for better understanding of biological mechanisms behind the disease. More recently, scientists have found that influences of disease risk factors can vary across disease subtypes by molecular pathological epidemiology (MPE) research (Ogino et al 2016). For example, the association of some life-style factors with the risk of colorectal cancer (CRC) is different across CRC subtypes (Ogino and Stampfer 2010; Campbell et al 2010). Statistical methodology has been developed to evaluate this type of etiologic heterogeneity across disease subtypes (Chatterjee et al 2010; Begg et al 2013; Wang et al 2015, 2016; Nevo et al 2016).

However, subtype data are often missing for some of the disease cases due to various reasons. For tumor tissue data, for example, biomarkers defining the tumor subtypes may be missing because tumor tissue is unavailable, or because collected tumor tissue is insufficient in quantity or quality. One commonly-used approach for estimation of the parameters in the proportional hazard (PH) competing risks models with missing cause of failure, originally suggested by Goetghebeur and Ryan (1995), is to construct estimating equations emerging from two separate partial likelihoods (Goetghebeur and Ryan 1995; Dewanji 1992). These two partial likelihood approaches were compared and carefully studied by Lu and Tsiatis (2005). More recently, Chatterjee et al (2010) suggested a two-stage model for the scenario where a large number of subtypes are composed from multiple markers, and generalized the estimating equation method of Goetghebeur and Ryan (1995) to estimate regression parameters under their two-stage model.

For this missing cause of failure problem, other methods developed include fully parametric methods (Flehinger et al 2002), Bayesian methods (Basu et al 2003; Sen et al 2010), as well as implementation of commonly used missing data methodology, such as multiple imputation (MI) (Lu and Tsiatis 2001), utilization of the EM algorithm (Craiu and Duchesne 2004), and inverse probability weighting (IPW) and augmented IPW (AIPW) approaches (Gao and Tsiatis 2005; Hyun et al 2012). Other competing risks models, such as additive hazards model (Lu and Liang 2008) or models for the cumulative incidence function (Bakoyannis et al 2010; Moreno-Betancur and Latouche 2013) were also investigated under missing cause of failure. Some of the aforementioned methods assume a missing-at-random (MAR) assumption that, given the event time and the covariates measured for all subjects, including subjects with and without the disease of interest, the probability of a missing cause is the same for all cases.

However, often, this MAR assumption is invalid. For example, it is possible that data on biomarkers, and therefore subtypes, are more likely to be missing when a tumor exhibits certain characteristics. In our motivating example, where of interest are the associations of risk factors with CRC subtypes microsatellite instability (MSI) and microsatellite stability (MSS) (Cancer Genome Atlas Network 2012; Kocarnik et al 2015), the MSI/MSS status is missing for more than half of the cases. Our data analysis suggested that CRC tumor

location (proximal colon, distal colon or rectum), which is measured for all cases we considered, may be indicative of the existence of adequate tumor tissue for obtaining the subtype data. Furthermore, it is well established in the subject-matter literature that CRC tumor location is associated with MSI/MSS status (Colussi et al 2013). In short, the goal of this paper is to provide methodology taking advantage of the information about the subtype status contained in some disease characteristics (for example, tumor location in our motivating example).

Existing methods allow for inclusion of auxiliary covariates in the weights for IPW or in the imputation model (Lu and Tsiatis 2001; Bakoyannis et al 2010). Gao and Tsiatis (2005) further developed an AIPW method, which is doubly-robust and yields consistent estimates when either the missingness probability or the imputation model are correctly specified. Here we present an alternative, which relies on a model for the distribution of the auxiliary covariates conditionally on the subtype and possibly time and other pre-diagnosis covariates. Our approach targets the exposure-disease associations across the different subtypes.

The rest of the paper is organized as follows. In Section 2, we present PH models for competing risks, allowing for missing cause of failure, and discuss various assumptions concerning the missingness pattern. In Section 3, we describe our approach in the presence of auxiliary case covariates which is valid under a weaker MAR assumption. Finite sample properties of the suggested method are investigated using simulations in Section 4. Analysis of our motivating example, the CRC subtype data, is presented in Section 5. Concluding remarks and discussion of available methods that allow for auxiliary covariates are given in Section 6.

2 Models and notations

We consider time-to-event data within the competing risks framework. Let $\tilde{T}$ be the time until the occurrence of the disease and Y be the subtype. Y takes a single value from $\{1, \dots, K\}$. Denote also $\mathbf{X} = \mathbf{X}(t)$ for a vector of possibly time-dependent covariates. A Cox-type model for the subtype-specific hazard is then

$$\lambda_k(t) := \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t \leq \tilde{T} < t + \Delta t, Y = k | \tilde{T} \geq t, \mathbf{X} = \mathbf{x}) = \lambda_{01}(t) \alpha_k(t; \boldsymbol{\eta}_k) \exp(\boldsymbol{\beta}_k^T \mathbf{x}) \quad (1)$$

with $\lambda_{01}(t)$ being an unspecified baseline hazard function for subtype 1 and where $\alpha_k(t; \boldsymbol{\eta}_k) = \lambda_{0k}(t)/\lambda_{01}(t)$, $k = 1, \dots, K$, are baseline hazard ratio functions; $\alpha_1(t) = 1$ by definition. This model allows an unspecified baseline hazard, as in the usual Cox model (Cox 1972), while allowing the baseline hazard ratio between two subtypes to depend on t , possibly via a parametric model. If the number of different subtypes is not too large, a possible approach is to specify a piecewise constant function for $\alpha_k(t; \boldsymbol{\eta}_k)$, for each k . While we operate under the commonly used proportional hazard model, the methodology we present can be extended to a broader set of regression functions. Let C be the censoring time, $T = \min(\tilde{T}, C)$, the observed time, and $\delta = I\{T = \tilde{T}\}$, the disease censoring indicator. We assume that given covariates, censoring time is independent of both $\tilde{T}$ and Y (Kalbeisch and Prentice 2011).

The model described above is extended to allow missing subtype data by introducing an indicator O that equals to one if the subtype Y is observed, and zero otherwise. If $\delta = 0$, then $O = 0$ as well. Denote

$$\pi(t, \mathbf{x}, k) := P(O = 1 | \tilde{T} = t, \mathbf{X} = \mathbf{x}, Y = k) \quad (2)$$

for the probability of observing the subtype given a diagnosis at time t , for a subject with a covariates vector $\mathbf{x}$ and subtype k . Goetghebeur and Ryan (1995) considered a model where the missing subtype probability depends on the event time ($\tilde{T}$) only. That is, $\pi(t, \mathbf{x}, k) = \pi(t) = P(O = 1 | \tilde{T} = t)$. This is a form of a MAR assumption that we denote by $\text{MAR}_{\tilde{T}}$. Similarly, we denote $\text{MAR}_{\tilde{T}, \mathbf{X}}$ for the commonly used assumption $\pi(t, \mathbf{x}, k) = \pi(t, \mathbf{x}) = P(O = 1 | \tilde{T} = t, \mathbf{X} = \mathbf{x})$. We also denote $\bar{\pi} = 1 - \pi$ as the probability of a missing subtype.

Often, in practice, subtypes (i.e., causes) are more likely to be missing when the disease exhibits specific characteristics which are associated with the subtypes, even after controlling for $\mathbf{X}$ and $\tilde{T}$. In these situations, the missingness probability depends on the subtype even after controlling for $\mathbf{X}$ and $\tilde{T}$; that is, neither $\text{MAR}_{\tilde{T}, \mathbf{X}}$ nor $\text{MAR}_{\tilde{T}}$ are expected to hold. However, often, the missing probability is independent of the subtype status when conditioning on these disease characteristics mentioned above in addition to $\mathbf{X}$ and $\tilde{T}$; this is a weaker assumption than $\text{MAR}_{\tilde{T}, \mathbf{X}}$ and $\text{MAR}_{\tilde{T}}$. We denote by $\mathbf{Q}$ the vector of measured disease characteristics, which we assume is available for all cases, and we denote its joint probability function by

$$\nu_k(\mathbf{q}, \mathbf{x}, t; \boldsymbol{\psi}_k) = P(\mathbf{Q} = \mathbf{q} | Y = k, \mathbf{X} = \mathbf{x}, \tilde{T} = t; \boldsymbol{\psi}_k),$$

with $\boldsymbol{\psi}_k$ being an unknown vector of parameters. In practice, $\mathbf{Q}$ might be conditionally independent of either $\tilde{T}$, $\mathbf{X}$ or both, given Y . Note that $\mathbf{Q}$ may include both discrete and continuous variables. It can also be a single random variable, as in the example presented in Section 5. In that example, $\mathbf{Q}$ is the tumor location and it is possibly associated with the probability of missing CRC subtype; see Section 5.

To account for $\mathbf{Q}$ effect on the possibility of missing subtype, the probability of observing the subtype, previously defined in (2), is redefined as

$$\pi(t, \mathbf{x}, \mathbf{q}, k) := P(O = 1 | \tilde{T} = t, \mathbf{X} = \mathbf{x}, \mathbf{Q} = \mathbf{q}, Y = k), \quad (3)$$

which is written as $\pi(t, \mathbf{x}, \mathbf{q}, k; \boldsymbol{\gamma})$ when a model with a vector of parameters $\boldsymbol{\gamma}$ is assumed. We can now introduce more relevant possible assumptions regarding the missing mechanism:

$$\begin{aligned} \pi(t, \mathbf{x}, \mathbf{q}, k) &= \pi(t, \mathbf{x}, \mathbf{q}) = P(O = 1 | \tilde{T} = t, \mathbf{X} = \mathbf{x}, \mathbf{Q} = \mathbf{q}), & (\text{MAR}_{\tilde{T}, \mathbf{X}, \mathbf{Q}}) \\ \pi(t, \mathbf{x}, \mathbf{q}, k) &= \pi(\mathbf{q}) = P(O = 1 | \mathbf{Q} = \mathbf{q}). & (\text{MAR}_{\mathbf{Q}}) \end{aligned}$$

Clearly, $MAR_{T,X,Q}$ is a weaker assumption than $MAR_{T,X}$. While subject-matter knowledge can help to determine which missing assumption is reasonable in a specific application, these assumptions cannot be tested using the observed data, as often the case when dealing with assumptions regarding missing data.

Let $W = (\tilde{T}, X, Q)$. When there are only two subtypes, the $MAR_{T,X,Q}$ assumption is equivalent to the assumption of Lu and Tsiatis (2001), which is, in our notation,

$$P(Y = 2|O = 1, W = w) = P(Y = 2|O = 0, W = w) = P(Y = 2|W = w). \quad (4)$$

Equivalence of $MAR_{T,X,Q}$ and (4) can be shown by

$$P(Y = 2|O = 1, W = w) = \frac{P(O = 1|Y = 2, W = w)P(Y = 2|W = w)}{P(O = 1|W = w)} = P(Y = 2|W = w),$$

where the last equality is justified by $MAR_{T,X,Q}$.

When the PH model holds, the validity of analysis conducted using existing methods relies on the missing-at-random assumption and correctness of a parametric model: $MAR_{T,X}$ and an accurate specification of $\pi(t, x; \gamma)$ (Goetghebeur and Ryan 1995); $MAR_{T,X,Q}$ and correct parametric specification of Y conditionally on $(X, \tilde{T}, Q)$ (Lu and Tsiatis 2001); or, at the very least, $MAR_{T,X,Q}$ and correct parametric specification of Y conditionally on $(X, \tilde{T}, Q)$ or of $\pi(t, x, q; \gamma)$ (Gao and Tsiatis 2005). Under $MAR_{T,X,Q}$, we present in the next section methodology that alleviate the concerns about the aforementioned parametric models while replacing them with assumptions on the conditional distribution of Q given $Y, \tilde{T}$ and X .

3 Estimation and inference under weaker MAR

We now move to construction of a partial likelihood for estimation of the model parameters $\eta = \{\eta_2, \dots, \eta_K\}$, $\psi = \{\psi_1, \dots, \psi_K\}$ and $\beta = \{\beta_1, \dots, \beta_k\}$, although β is of our main concern. By separating joint distributions into conditional and marginal distributions, we may now write under $MAR_{T,X,Q}$ (for all t, x and q)

$$\begin{aligned} & \lim_{\Delta t \downarrow 0} P(t \leq \tilde{T} < t + \Delta t, Y = k, Q = q, O = 1 | \tilde{T} \geq t, X = x) \\ &= P(O = 1 | \tilde{T} = t, X = x, Q = q) \times P(Q = q | Y = k, \tilde{T} = t, X = x) \\ &\times \lim_{\Delta t \downarrow 0} P(t \leq \tilde{T} < t + \Delta t, Y = k | \tilde{T} \geq t, X = x) = \pi(t, x, q) \nu_k(q, x, t; \psi_k) \lambda_{01}(t) \alpha_k(t; \eta_k) \exp \\ &(\beta_k^T x) \end{aligned} \quad (5)$$

if the subtype Y is observed, and

$$\begin{aligned}
& \lim_{\Delta t \downarrow 0} P(t \leq \tilde{T} < t + \Delta t, \mathbf{Q} = \mathbf{q}, O = 0 | \tilde{T} \geq t, \mathbf{X} = \mathbf{x}) \\
&= \sum_{k=1}^K \lim_{\Delta t \downarrow 0} P(t \leq \tilde{T} < t + \Delta t, Y = k, \mathbf{Q} = \mathbf{q}, O = 0 | \tilde{T} \geq t, \mathbf{X} = \mathbf{x}) \\
&= \sum_{k=1}^K \bar{\pi}(t, \mathbf{x}, \mathbf{q}) \nu_k(\mathbf{q}, \mathbf{x}, t; \boldsymbol{\psi}_k) \lambda_{01}(t) \alpha_k(t; \boldsymbol{\eta}_k) \exp(\boldsymbol{\beta}_k^T \mathbf{x})
\end{aligned} \tag{6}$$

if the subtype is not observed.

Construction of partial likelihood for the PH competing risks model with missing subtypes for part of the cases entails some complexity. To present existing methods, we first consider the situation where $\mathbf{Q}$ is not observed and is not part of the model. For the PH competing risks model, the standard approach when subtypes are observed for all cases is to condition not only on the case occurrence, as in the standard Cox model with a single type of event, but also on its subtype. Then, in a model without $\mathbf{Q}$, the baseline hazard in the denominator and the numerator cancel out, $\pi_i = 1$ for all i , and the resulting partial likelihood for $\boldsymbol{\beta}$ in the PH competing risks model is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \left[\frac{\exp(\boldsymbol{\beta}_{y_i}^T \mathbf{x}_i)}{\sum_{j=1}^n \xi_j(t_i) \exp(\boldsymbol{\beta}_{y_j}^T \mathbf{x}_j)} \right]^{I\{\delta_i = 1\}}$$

(Prentice et al 1978), where $\xi_j(t)$ equals to one if j is in the risk set at time t and zero otherwise and y_i is the observed subtype.

When subtypes are unobserved for some cases, two partial likelihood approaches, which differ in the conditioned event, have been discussed, typically under $\text{MAR}_{T, \mathbf{X}}$, or stronger assumptions. The first (Goetghebeur and Ryan 1995) is to condition on the occurrence of an event, the observed subtype and on the fact that subtype was observed, and for events with unobserved subtype, to condition on the occurrence of an event and on not observing the corresponding subtype. This is done by considering the hazard at time t of event occurrence of any subtype, which is the sum of the hazard functions for all subtypes, also at time t . That is,

$$P(t \leq \tilde{T} < t + \Delta t | \tilde{T} \geq t, \mathbf{X} = \mathbf{x}) = \sum_{k=1}^K P(t \leq \tilde{T} < t + \Delta t, Y = k | \tilde{T} \geq t, \mathbf{X} = \mathbf{x}).$$

Then, incorporating the missingness model under $\text{MAR}_{T, \mathbf{X}}$ and by using calculations similar to (5) and (6), this likelihood (Goetghebeur and Ryan 1995) becomes

$$L(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) = \left\{ \prod_{i=1}^n \left[\frac{\pi(t_i, \mathbf{x}_i; \boldsymbol{\gamma}) \exp(\boldsymbol{\beta}_{y_i}^T \mathbf{x}_i)}{\sum_{j=1}^n \xi_j(t_i) \pi(t_i, \mathbf{x}_j; \boldsymbol{\gamma}) \exp(\boldsymbol{\beta}_{y_i}^T \mathbf{x}_j)} \right]^{I\{\delta_i=1, O_i=1\}} \right. \\ \left. \times \left[\frac{\bar{\pi}(t_i, \mathbf{x}_j; \boldsymbol{\gamma}) \sum_{k=1}^K \alpha_k(t_i, \boldsymbol{\eta}_k) \exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)}{\sum_{j=1}^n \xi_j(t_i) \bar{\pi}(t_i, \mathbf{x}_j; \boldsymbol{\gamma}) \sum_{m=1}^K \alpha_m(t_i, \boldsymbol{\eta}_m) \exp(\boldsymbol{\beta}_m^T \mathbf{x}_j)} \right]^{I\{\delta_i=1, O_i=0\}} \right\}. \quad (7)$$

However, when $\text{MAR}_{T, X}$ does not hold, this likelihood cannot be used under the weaker $\text{MAR}_{T, X, Q}$ assumption since π could not be calculated even when Q is observed. This is because, under $\text{MAR}_{T, X, Q}$, π depends on Q , which is only observed for subjects in the risk set with $\delta = 1$, and thus π cannot be calculated for members of the risk set with $\delta = 0$.

We now construct a more informative partial likelihood, similar to the one suggested by Dewanji (1992). We condition on the fact that event was observed, but not on the event subtype. Taking this approach, we can write down the partial likelihood for the observed data $(T_i, \delta_i, X_i, Y_i O_i, O_i, Q_i \delta_i)$. For censored observations, O_i is defined as zero and Y_i and Q_i are taken to be missing. Following (1), (5) and (6) and assuming $\text{MAR}_{T, X, Q}$, we have

$$L_Q^*(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\psi}) = \prod_{i=1}^n \left\{ \left[\frac{\pi(t_i, \mathbf{x}_i, \mathbf{q}_i; \boldsymbol{\gamma}) \nu_{y_i}(\mathbf{q}, \mathbf{x}, t; \boldsymbol{\psi}_{y_i}) \alpha_{y_i}(t_i; \boldsymbol{\eta}_{y_i}) \exp(\boldsymbol{\beta}_{y_i}^T \mathbf{x}_i)}{\sum_{j=1}^n \xi_j(t_i) \sum_{m=1}^K \alpha_m(t_i; \boldsymbol{\eta}_m) \exp(\boldsymbol{\beta}_m^T \mathbf{x}_j)} \right]^{I\{\delta_i=1, O_i=1\}} \right. \\ \left. \times \left[\frac{\sum_{k=1}^K \bar{\pi}(t_i, \mathbf{x}_i, \mathbf{q}_i; \boldsymbol{\gamma}) \nu_k(\mathbf{q}, \mathbf{x}, t; \boldsymbol{\psi}_k) \alpha_k(t_i; \boldsymbol{\eta}_k) \exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)}{\sum_{j=1}^n \xi_j(t_i) \sum_{m=1}^K \alpha_m(t_i; \boldsymbol{\eta}_m) \exp(\boldsymbol{\beta}_m^T \mathbf{x}_j)} \right]^{I\{\delta_i=1, O_i=0\}} \right\}. \quad (8)$$

Note that when using L_Q^* , the missing model need not to be specified, as long as at least $\text{MAR}_{T, X, Q}$ holds. This is because we can decompose L_Q^* into $L_{Q1}^* \times L_{Q2}^*$, where

$$L_{Q1}^*(\boldsymbol{\gamma}) = \prod_{i=1}^n \left[\left(\pi(t_i, \mathbf{x}_i, \mathbf{q}_i; \boldsymbol{\gamma}) \right)^{I\{\delta_i=1, O_i=1\}} \bar{\pi}(t_i, \mathbf{x}_i, \mathbf{q}_i; \boldsymbol{\gamma})^{I\{\delta_i=1, O_i=0\}} \right],$$

$$L_{Q2}^*(\beta, \eta) = \prod_{i=1}^n \left\{ \left[\frac{\nu_{y_i}(\mathbf{q}, \mathbf{x}, t; \boldsymbol{\psi}_{y_i}) \alpha_{y_i}(t_i; \boldsymbol{\eta}_{y_i}) \exp(\boldsymbol{\beta}_{y_i}^T \mathbf{x}_i)}{\sum_{j=1}^n \xi_j(t_i) \sum_{m=1}^K \alpha_m(t_i; \boldsymbol{\eta}_m) \exp(\boldsymbol{\beta}_m^T \mathbf{x}_j)} \right]^{I\{\delta_i=1, O_i=1\}} \right. \\ \left. \times \left[\frac{\sum_{k=1}^K \nu_k(\mathbf{q}, \mathbf{x}, t; \boldsymbol{\psi}_k) \alpha_k(t_i; \boldsymbol{\eta}_k) \exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)}{\sum_{j=1}^n \xi_j(t_i) \sum_{m=1}^K \alpha_m(t_i; \boldsymbol{\eta}_m) \exp(\boldsymbol{\beta}_m^T \mathbf{x}_j)} \right]^{I\{\delta_i=1, O_i=0\}} \right\}. \quad (9)$$

Estimation of β can be done based solely on L_{Q2}^* . The approach of Goetghebeur and Ryan (1995), which is also utilized in Chatterjee et al (2010), is to combine the score function for β from L in (7) and the score function for η from a likelihood L^* , a likelihood analogue of L_Q^* in (8) for a model without Q , to form a system of estimating equations. However, as we noted before, L cannot be used when the missing probability depends on Q . It is worth mentioning that $\nu_k(\mathbf{q}, \mathbf{x}, t; \boldsymbol{\psi}_k)$ cannot be estimated from the observed subtype data only since, if O and Q are conditionally dependent,

$$\nu_k(\mathbf{q}, \mathbf{x}, t; \boldsymbol{\psi}_k) = P(Q|X, T; \boldsymbol{\psi}_k) \neq P(Q|O=1, X, T; \boldsymbol{\psi}_k, \gamma).$$

Let $\theta = (\beta, \eta, \psi)$ and let θ^0 be its true value. We suggest to estimate θ by maximizing L_{Q2}^* (or $\ell_{Q2}^* = \log L_{Q2}^*$). This could be done using a standard Newton-Raphson algorithm. Denote $\hat{\theta}$ for the estimated vector of parameters. It can be shown that $\hat{\theta}$ is a consistent estimator. Furthermore, $\sqrt{n}(\hat{\theta} - \theta^0)$ converges in distribution to a multivariate normal distribution with a covariance matrix that can be consistently estimated by a sandwich estimator. Details and proofs are given in Web Appendix A.

4 Simulation Study

In the main simulation study, we considered two possible subtypes. The baseline hazard for each subtype was of a Weibull distribution, with parameter values chosen such that the baseline hazard ratio equals to $\alpha_2(t; \eta_1, \eta_2) = \eta_1 t^{\eta_2}$ with $\eta_1 \approx 0.037$ and $\eta_2 = 1$. We took X to be a single binary covariate with a prevalence of 0.4 and let the corresponding hazard ratios to be 1.25 and 2.5 in subtypes 1 and 2, respectively. Censoring was generated using an exponential distribution with mean 50, with additional type I censoring at time 90. The above parameter values were chosen to have a censoring rate of about 70%. We took Q , the auxiliary variable, to be a single binary variable that we generated according to $P(Q=1|Y=1) = 0.25$ and $P(Q=1|Y=2) = 0.5$.

We first considered the MAR_Q scenario under a logistic regression model for the missing $\pi(q) = \text{expit}(\gamma_0 + \gamma_q q)$. We considered various values for γ_0 and γ_q determined by our desired values for $P(O=1|Q=0)$ and $P(O=1|Q=1)$. In this scenario, we compared between three alternatives. The first was a complete case analysis (CCA), where cases with

unknown subtype were thrown away and a standard competing risks model was used. The second was our method, where estimates were obtained by maximization of L_{Q2}^* in (9), and variance was estimated using the asymptotic variance estimator given in Web Appendix A. The third alternative was the MI method of Lu and Tsiatis (2001), with 10 imputations. While data were simulated under MAR_Q , and the distribution of Q depends solely on Y , a correctly specified imputation model of Y is given as a logistic regression of Y with Q , $\log(\tilde{T})$ and X as the covariates. This can be verified by Bayes theorem; See Web Appendix B for the details. Therefore, we included Q , X and $\log(\tilde{T})$ as covariates in the imputation model. The variance estimator given by Lu and Tsiatis (2001) was used.

Summary of 1000 simulation iterations is presented in Table 1 for sample sizes $n = 1000$, 10000 with approximately 300 and 3000 numbers of cases, respectively. We report relative bias (in percentage), mean estimated standard error, empirical standard deviation and coverage rates of confidence intervals when using our methods. Confidence intervals were constructed using the asymptotic normal distribution. While our main motivation was cancer incidence studies with large sample size and high censoring rate, we consider also studies with moderate sample size ($n = 500$) and different censoring rates (30%, 50% and 70%). Table 2 presents these results for β_1 . The results for β_2 are presented in Web Appendix C (Web Table 1). Both our method and the MI approach showed good finite sample properties, with desirable empirical coverage rates for the corresponding confidence intervals. There was a large bias in the estimation by CCA. Compared to MI, L_Q^* was almost always slightly more efficient.

We performed additional simulations to evaluate sensitivity of the methods to misspecification of the baseline hazard ratio. We considered the Gompertz distribution for the baseline hazard instead of the Weibull, under moderate sample size and varying censoring rates. The analysis was conducted assuming $\alpha_2(t; \eta_1, \eta_2) = \eta_1 t^{\eta_2}$, even though this model does not hold for the Gompertz distribution. Therefore, an increase in the bias of our method was expected, due to misspecification of the baseline hazard ratio. As reported in Web Appendix C (Web Tables 2 and 3), a larger bias was observed under this scenario, and the bias of our method and of the MI method was similar.

We considered an additional the sensitivity analysis, now for the assumption that the distribution of Q depends on Y . We simulated Q as before for $X = 0$, i.e., $P(Q = 1|X = 0, Y = 1) = 0.25$ and $P(Q = 1|X = 0, Y = 2) = 0.5$, but for $X = 1$, we simulated Q according to $P(Q = 1|X = 1, Y = 1) = 0.35$ and $P(Q = 1|X = 1, Y = 2) = 0.6$. In practice, we could have included X as part of the model for the distribution of Q , as we do in our motivating example, where we include a risk factor in the distribution of Q . However, the goal of this simulation was to study the effect of misspecification of the distribution of Q on L_Q^* . The imputation model used by MI was the same as before (i.e., including X in linear form). This model is misspecified because the functional form of X in the imputation model is complex and non-linear; this can be shown using the same approach as in Web Appendix B. The results for $n = 1000$, 10000 are presented in Web Appendix C (Web Table 4). For some scenarios the bias

was larger comparing to Table 1, especially for β_1 . The standard deviation of the two leading methods (L_Q^* and MI) increased.

We then turned to more complex missingness models. First, we simulated the data under $MAR_{T,X,Q}$ with the model $\pi(t, x, q) = \text{expit}(\gamma_q q + 0.5x - 0.01I\{t > 50\})$, with values $\log(\gamma_q) \in \{1, 1.25, 1.75, 2.5, 5\}$. Larger γ_q value implies less cases with missing subtypes, and a more substantial role of Q in determining O . Other than the missing model, we used the same values as in the MAR_Q scenario. In addition to CCA, MI and our method, we also considered the estimating equations of Goetghebeur and Ryan (1995) (GR), which is valid under $MAR_{T,X}$. For GR, variance estimates and confidence intervals were obtained by the bootstrap with 500 repetitions. As a final scenario, we repeated the aforementioned simulations, but now taking the not-missing-at-random (NMAR) model

$$\pi(t, x, k) = P(O = 1 | \tilde{T} = t, X = x, Y = k) = \text{expit}(\gamma_y I\{k = 2\} - 0.01I\{t > 50\} + 0.5x)$$

to be the correct model. That is, now all methods compared are theoretically invalid. We considered $\log(\gamma_y) \in \{1.25, 1.75, 2.5\}$.

Tables 3 and 4 present the results for $MAR_{T,X,Q}$ and NMAR scenarios, respectively, for $n = 1,000$. While larger γ_q or γ_y values imply less missing subtypes, larger bias (in absolute value) was observed for estimates obtained by the GR method. For $MAR_{T,X,Q}$, performance of the L_Q^* approach was not affected by the value of γ_q . Looking at the results for NMAR, L_Q^* and MI had larger bias compared with $MAR_{T,X,Q}$ scenario, especially for β_2 . The bias of L_Q^* and MI was substantially lower than the bias observed for CCA and GR. Throughout all simulations scenarios, our method remained the most efficient.

The likelihood suggested by Dewanji (1992) had been criticized for having a non mean-zero score function for β if α_k is misspecified. To investigate this potential limitation, we considered the L_Q^* estimator when taking a piecewise constant function for the baseline hazard ratio α_k and the data was generated by the same model outlined above. The results, presented in Web Appendix C (Web Table 5), did not reveal further bias when using piecewise constant function for the baseline hazard ratio α_k .

5 Colorectal cancer analysis

To illustrate the use of our method, we present in this section an analysis of CRC data. CRC is regularly classified according to results of molecular studies (Ogino et al 2011; Ogino and Stampfer 2010; Campbell et al 2010; Kuipers et al 2015). One commonly-used classification is microsatellite instability (MSI) versus microsatellite stability (MSS). For example, high body mass index (BMI) has been associated with MSS subtype of CRC tumors but not MSI subtype (Hughes et al 2012).

We used the Nurses' Health Study (NHS) cohort, in which 121,701 female nurses enrolled in 1976 and since then answered biennial questionnaire regarding lifestyle and other risk

factors. Whenever a new CRC cancer was diagnosed, the date was recorded as well as basic tumor characteristics, available for all cases we considered. The tumor location is classified to be proximal colon, distal colon or rectum. Our analysis included 1,844 CRC cases that were observed, of which 121 MSI cases, 477 MSS cases, and the remaining cases (66%) had missing MSI status. In total, our data include 114,073 participants, with 2,786,825 person-years of follow-up. Data on some risk factors were not measured from the beginning of the study, hence not all of the 121,701 participants were included in the data.

As a preliminary analysis, we considered a logistic regression model for O , the indicator that MSI status was known, and we examined various covariates. Details are given in Web Appendix D. As presented in Web Appendix D, missing MSI status depends on the tumor location, even when including other covariates (X) and the time of the diagnosis (T). It was previously established that in CRC, MSI status and proximal location are associated (Colussi et al 2013). Thus, the traditional $MAR_{T,X}$ probably does not hold here. We consider the following risk factors: family history of CRC (binary), BMI (continuous), regular aspirin use (binary) and cumulative pack-year of smoking (continuous). We take here Q to be the tumor location, equals to one for proximal tumors and zero otherwise. Based on subject-matter considerations, we include in the model for the CRC tumor location the subtype MSI status (Y), and the continuous covariate BMI. That is, we let $P(Q=1|Y=k, BMI) = \text{expit}(\psi_0 + \psi_1 I\{k=2\} + \psi_2 BMI)$. To account for changes in the baseline hazard along the (calendar) years, we used a stratified version for the baseline hazard ratio α_2 . The stratified version, presented in detail in Web Appendix E, is a natural generalization of L_Q^* given by (8).

Table 5 compares the results between a CCA with a stratified Cox model for the subtype-specific hazards, with the 598 cases with known MSI status and all CRC-free participants, and implementation of our method based on L_Q^* that uses data on all 1844 cases and all CRC-free participants. Differences are shown between the estimates in both methods. For example, for CRC MSI subtype the log-hazard ratio of aspirin is about 70% larger (in absolute value) when using CCA comparing to using L_Q^* . The standard errors of the L_Q^* -estimates are dramatically lower than those obtained by CCA. Furthermore, the association of aspirin with MSS CRC incidence that was not significant at 5% level when using CCA is significant when using L_Q^* .

6 Discussion

Some existing methods for subtype analysis rely on missing-at-random assumptions that may fail to hold in practice. Analysis under these assumptions potentially results in biased estimates. In this paper, we presented a weaker missing-at-random assumption which is reasonable when additionally auxiliary covariates are measured for all cases. Under this assumption, which implies and implied by the assumption of Lu and Tsiatis (2001), we have developed an informative partial likelihood approach which results in consistent and normally distributed estimators.

Alternatives to our L_Q^* method, MI (Lu and Tsiatis 2001), IPW, or AIPW (Gao and Tsiatis 2005), can be used under $MAR_{T,X,Q}$. The MI approach relies on correct specification of the distribution of Y conditionally on the time, risk-factors and auxiliary covariates. Furthermore, if a PH model is assumed for the cause-specific hazard functions in all subtypes, correct specification of the imputation model implicitly depend on the true baseline hazard functions and the conditional distribution of Q (Web Appendix B). IPW relies on correct specification of the missing mechanism. A doubly-robust AIPW method by Gao and Tsiatis (2005) is valid if only one of the two aforementioned models is correctly specified. Our method does not require correct specification of any of these models. Instead, it is valid whenever the model for Q conditionally on the subtype, risk-factors and time is correctly specified. Our method is especially attractive when researchers have prior knowledge on the relationship between the auxiliary covariates and the subtype, as in our motivating example, where the association between tumor location and MSI was well-established.

Our simulation results suggest an efficiency gain comparing to the MI method when the MI and conditional distribution of Q models are correctly specified. However, in practice, the models assumed by the two methods are not necessarily simultaneously correctly specified, as a simple model in one method may imply a complicated, or non-standard, model for the other method. Therefore, it is recommended to use the method that subject-matter can better direct us what model to construct. As shown in the simulation study, misspecification of the distribution of Q , specifically by neglecting to include certain covariate, would potentially result in biased estimates.

If Q is high-dimensional, modeling the distribution of Q may increase the total number of model parameters. It worth mentioning that from our point of view, $\boldsymbol{\psi}$, the parameters characterizing the distribution of Q are nuisance parameters, and exact correct specification of the joint distribution of Q is not our goal. Thus, in order to get a low-dimensional vector or even a scalar Q to be used in the main analysis, a preliminary analysis of Q yielding a working model to summarize Q into low-dimensional vector (or scalar) $h(Q)$ can be used. For example, one may apply a principal component analysis or a factor analysis. Note that, even in this high-dimensional case, $\boldsymbol{\pi}$ is left unspecified so problems typically arising when missing data depends on high-dimensional variables are circumvented.

When the number of subtypes of interest is relatively small, which is generally the case, using a piecewise constant function for the baseline hazard ratio $\alpha_k(t, \boldsymbol{\eta}_k)$ can be a good alternative to a more restrictive parametric model. If the number of investigated subtypes is very large, our approach may result in a computational burden when estimating the large number of parameters. Under the more restrictive $MAR_{T,X}$, the method suggested by Chatterjee et al (2010) may serve as a potential good alternative. In their suggested method, the auxiliary covariates can be used as an additional biomarker, or “disease trait” in their terminology, that its crossproduct with other disease traits form the subtypes. From a clinical point of view, however, biomarkers (such as MSI) are usually used for subtype definition, while the other characteristics (such as tumor location) are treated as specific clinical properties of the disease.

R code implementing our methods is available upon request. Analysis using our approach will result in two major advantages for practitioners. First, potential bias will be eliminated. Second, data that are not always used today can be utilized to get more efficient estimates, with better power for hypothesis testing and narrower confidence intervals as a result of inclusion of cases that had been left aside so far. Hence our methods can be very useful in the era of MPE and precision medicine.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank two anonymous reviewers and the associate editor for insightful comments and suggestions that improved the paper.

This research was funded by NIH grants R35 CA197735, R01 CA151993, K07 CA190673, P50 CA127003, UMI CA186107 and P01 CA87969.

References

- Bakoyannis G, Siannis F, Touloumi G. Modelling competing risks data with missing cause of failure. *Statistics in medicine*. 2010; 29(30):3172–3185. [PubMed: 21170911]
- Basu S, Sen A, Banerjee M. Bayesian analysis of competing risks with partially masked cause of failure. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2003; 52(1):77–93.
- Begg CB, Zabor EC, Bernstein JL, Bernstein L, Press MF, Seshan VE. A conceptual and methodological framework for investigating etiologic heterogeneity. *Statistics in medicine*. 2013; 32(29):5039–5052. [PubMed: 23857589]
- Campbell PT, Jacobs ET, Ulrich CM, Figueiredo JC, Poynter JN, McLaughlin JR, Haile RW, Jacobs EJ, Newcomb PA, Potter JD, et al. Case–control study of overweight, obesity, and colorectal cancer risk, overall and by tumor microsatellite instability status. *Journal of the National Cancer Institute*. 2010
- Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012; 487(7407):330–337. [PubMed: 22810696]
- Chatterjee N, Sinha S, Diver WR, Feigelson HS. Analysis of cohort studies with multivariate and partially observed disease classification data. *Biometrika*. 2010; 97(3):683–698. [PubMed: 22822252]
- Colussi D, Brandi G, Bazzoli F, Ricciardiello L. Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention. *International journal of molecular sciences*. 2013; 14(8):16,365–16,385.
- Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society Series B (Methodological)*. 1972; 34(2):187–220.
- Craiu RV, Duchesne T. Inference based on the EM algorithm for the competing risks model with masked causes of failure. *Biometrika*. 2004; 91(3):543–558.
- Dewanji A. A note on a test for competing risks with missing failure type. *Biometrika*. 1992; 79(4): 855–857.
- Flehinger BJ, Reiser B, Yashchin E. Parametric modeling for survival with competing risks and masked failure causes. *Lifetime Data Analysis*. 2002; 8(2):177–203. [PubMed: 12048866]
- Gao G, Tsiatis AA. Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika*. 2005; 92(4):875–891.
- Goetghebuer E, Ryan L. Analysis of competing risks survival data when some failure types are missing. *Biometrika*. 1995; 82(4):821–833.

- Hughes LA, Williamson EJ, van Engeland M, Jenkins MA, Giles GG, Hopper JL, Southey MC, Young JP, Buchanan DD, Walsh MD, et al. Body size and risk for colorectal cancers showing braf mutations or microsatellite instability: a pooled analysis. *International journal of epidemiology*. 2012;dys055.
- Hyun S, Lee J, Sun Y. Proportional hazards model for competing risks data with missing cause of failure. *Journal of statistical planning and inference*. 2012; 142(7):1767–1779. [PubMed: 22468017]
- Kalbeisch, JD., Prentice, RL. The statistical analysis of failure time data. Vol. 360. John Wiley & Sons; 2011.
- Kocarnik JM, Shiovitz S, Phipps AI. Molecular phenotypes of colorectal cancer and potential clinical applications. *Gastroenterology report*. 2015:gov046.
- Kuipers EJ, Grady WM, Lieberman D, Seufferlein T, Sung JJ, Boelens PG, van de Velde CJH, Watanabe T. Colorectal cancer. *Nature Reviews Disease Primers*. 2015; 1:15,065. EP –, URL <http://dx.doi.org/10.1038/nrdp.2015.65>.
- Lu K, Tsiatis AA. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*. 2001; 57(4):1191–1197. [PubMed: 11764260]
- Lu K, Tsiatis AA. Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. *Lifetime Data Analysis*. 2005; 11(1):29–40. [PubMed: 15747588]
- Lu W, Liang Y. Analysis of competing risks data with missing cause of failure under additive hazards model. *Statistica Sinica*. 2008:219–234.
- Moreno-Betancur M, Latouche A. Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values. *Statistics in medicine*. 2013; 32(18):3206–3223. [PubMed: 23653257]
- Nevo D, Zucker DM, Tamimi RM, Wang M. Accounting for measurement error in biomarker data and misclassification of subtypes in the analysis of tumor data. *Statistics in Medicine*. 2016; 35(30): 5686–5700. [PubMed: 27558651]
- Ogino S, Stampfer M. Lifestyle factors and microsatellite instability in colorectal cancer: the evolving field of molecular pathological epidemiology. *Journal of the National Cancer Institute*. 2010; 102(6):365–367. [PubMed: 20208016]
- Ogino S, Chan AT, Fuchs CS, Giovannucci E. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. *Gut*. 2011; 60(3):397–411. [PubMed: 21036793]
- Ogino S, Nishihara R, VanderWeele TJ, Wang M, Nishi A, Lochhead P, Qian ZR, Zhang X, Wu K, Nan H, et al. The role of molecular pathological epidemiology in the study of neoplastic and non-neoplastic diseases in the era of precision medicine. *Epidemiology (Cambridge, Mass)*. 2016
- Prentice RL, Kalbeisch JD, Peterson AV Jr, Flournoy N, Farewell V, Breslow N. The analysis of failure times in the presence of competing risks. *Biometrics*. 1978; 43(4):541–554.
- Sen A, Banerjee M, Li Y, Noone AM. A Bayesian approach to competing risks analysis with masked cause of death. *Statistics in medicine*. 2010; 29(16):1681–1695. [PubMed: 20575048]
- Wang M, Kuchiba A, Ogino S. A meta-regression method for studying etiological heterogeneity across disease subtypes classified by multiple biomarkers. *American journal of epidemiology*. 2015:kww040.
- Wang M, Spiegelman D, Kuchiba A, Lochhead P, Kim S, Chan AT, Poole EM, Tamimi R, Tworoger SS, Giovannucci E, Rosner B, Ogino S. Statistical methods for studying disease subtype heterogeneity. *Statistics in Medicine*. 2016; 35(5):782–800. URL <http://dx.doi.org/10.1002/sim.6793>. DOI: 10.1002/sim.6793 [PubMed: 26619806]

Table 1

Simulation results under MAR_Q for complete case analysis (CCA), our method (L_Q^*) and multiple imputation (MI). True parameters values were $\beta_1 = 0.223 = \log(1.25)$ and $\beta_2 = 0.916 = \log(2.5)$. We present relative bias (%), empirical standard deviation (SD.EMP), mean estimated standard error (SE.EST) and confidence interval coverage rate with 95% confidence level (CP95).

	$P(O = 1 Q)$		%Missing	Subtype	$n = 1,000$				$n = 10,000$			
	$Q = 0$	$Q = 1$			CCA	L_Q^*	MI	CCA	L_Q^*	MI	CCA	L_Q^*
β_1	0.20	0.80	55%	Bias(%)	48.362	-3.942	-3.881	50.260	-0.902	-2.048		
				SD.EMP	0.348	0.267	0.287	0.106	0.082	0.085		
				SE.EST	0.338	0.265	0.275	0.105	0.081	0.085		
				CP95	0.933	0.950	0.946	0.819	0.944	0.950		
	0.40	0.80	44%	Bias(%)	33.678	-7.044	-8.073	35.171	-3.266	-3.552		
				SD.EMP	0.292	0.247	0.255	0.087	0.074	0.077		
				SE.EST	0.281	0.240	0.246	0.088	0.074	0.076		
				CP95	0.934	0.951	0.951	0.855	0.954	0.952		
	0.80	0.80	20%	Bias(%)	12.829	-5.603	-5.538	14.338	-2.943	-2.591		
				SD.EMP	0.227	0.215	0.216	0.069	0.067	0.067		
β_2	0.80	0.40	36%	SE.EST	0.221	0.211	0.212	0.069	0.066	0.066		
				CP95	0.945	0.952	0.952	0.928	0.938	0.946		
				Bias(%)	32.265	-7.567	-6.891	38.340	-0.009	0.313		
				SD.EMP	0.239	0.223	0.228	0.076	0.071	0.072		
	0.80	0.20	45%	SE.EST	0.237	0.222	0.224	0.074	0.069	0.070		
				CP95	0.941	0.957	0.951	0.774	0.945	0.942		
				Bias(%)	41.585	-8.641	-7.542	48.070	-2.394	-2.197		
				SD.EMP	0.251	0.231	0.237	0.076	0.071	0.074		
	0.80	0.80	55%	SE.EST	0.246	0.229	0.232	0.077	0.071	0.073		
				CP95	0.936	0.955	0.958	0.716	0.952	0.948		
β_2	0.20	0.80	55%	Bias(%)	15.535	1.124	1.865	13.291	-1.065	-0.205		
				SD.EMP	0.208	0.176	0.185	0.066	0.055	0.057		
				SE.EST	0.211	0.178	0.184	0.066	0.055	0.057		

$P(O = 1 Q)$		%Missing	$n = 1,000$			$n = 10,000$			
$Q = 0$	$Q = 1$	Subtype	CCA	L^*_Q	MI	CCA	L^*_Q	MI	
0.40	0.80	44%	CP95	0.909	0.949	0.941	0.545	0.944	0.953
			Bias(%)	10.724	-0.401	0.396	9.917	-1.026	-0.289
			SD.EMP	0.198	0.165	0.169	0.057	0.049	0.051
			SE.EST	0.192	0.166	0.169	0.060	0.052	0.053
			CP95	0.916	0.948	0.943	0.678	0.962	0.964
0.80	0.80	20%	Bias(%)	5.046	-0.161	0.499	4.645	-0.497	0.158
			SD.EMP	0.166	0.152	0.154	0.051	0.047	0.048
			SE.EST	0.166	0.154	0.155	0.052	0.048	0.048
			CP95	0.940	0.954	0.952	0.870	0.955	0.956
			Bias(%)	10.592	0.056	0.614	9.970	-0.622	0.019
0.80	0.40	36%	SD.EMP	0.187	0.155	0.157	0.060	0.049	0.050
			SE.EST	0.192	0.160	0.161	0.060	0.050	0.050
			CP95	0.941	0.962	0.961	0.687	0.952	0.953
			Bias(%)	12.730	-0.717	-0.222	13.232	-0.578	0.117
			SD.EMP	0.212	0.166	0.169	0.066	0.052	0.053
0.80	0.20	45%	SE.EST	0.211	0.165	0.166	0.066	0.051	0.052
			CP95	0.927	0.952	0.953	0.553	0.944	0.948

Table 2

Simulation results ($n = 500$) for different censoring rates under MAR_{Q_0} . Methods compared are complete case analysis (CCA), our method (L^*_Q) and multiple imputation (MI). True parameter values were $\beta_1 = \log(1.25)$ and $\beta_2 = \log(2.5)$. Results presented for β_1 . We present relative bias (%), empirical standard deviation (SD.EMP), mean estimated standard error (SE.EST) and confidence interval coverage rate with 95% confidence level (CP95).

$P(O = 1 Q)$	$Q = 0$	$Q = 1$	%Missing	Subtype	30% censoring			50% censoring			70% censoring		
					CCA	L^*_Q	MI	CCA	L^*_Q	MI	CCA	L^*_Q	MI
0.20	0.80	55%		Bias(%)	65.838	-15.338	-17.930	63.329	-13.382	-13.283	44.074	-13.438	-16.010
				SD.EMP	0.356	0.284	0.300	0.405	0.320	0.337	0.502	0.400	0.428
				SE.EST	0.343	0.279	0.290	0.395	0.317	0.330	0.495	0.392	0.406
				CP95	0.922	0.949	0.948	0.928	0.963	0.964	0.957	0.963	0.960
0.40	0.80	44%		Bias(%)	49.801	-7.809	-8.186	41.158	-9.539	-11.089	37.044	-3.834	-4.810
				SD.EMP	0.271	0.241	0.246	0.333	0.281	0.290	0.406	0.348	0.363
				SE.EST	0.283	0.246	0.252	0.327	0.283	0.289	0.404	0.347	0.353
				CP95	0.941	0.964	0.965	0.935	0.958	0.952	0.946	0.961	0.960
0.80	0.80	20%		Bias(%)	16.931	-8.812	-8.592	16.429	-6.220	-4.993	7.714	-7.944	-7.685
				SD.EMP	0.230	0.222	0.224	0.263	0.247	0.250	0.321	0.304	0.308
				SE.EST	0.222	0.213	0.214	0.256	0.245	0.246	0.316	0.300	0.302
				CP95	0.938	0.948	0.949	0.944	0.956	0.949	0.948	0.955	0.954
0.80	0.40	36%		Bias(%)	53.182	-6.668	-5.899	47.576	-3.253	-3.444	31.792	-5.150	-4.687
				SD.EMP	0.230	0.220	0.225	0.284	0.266	0.271	0.371	0.345	0.352
				SE.EST	0.238	0.226	0.228	0.275	0.260	0.262	0.341	0.321	0.324
				CP95	0.934	0.960	0.955	0.921	0.942	0.945	0.922	0.937	0.940
0.80	0.20	45%		Bias(%)	76.348	-4.108	-2.223	63.310	-4.677	-4.687	48.586	-2.863	-2.326
				SD.EMP	0.254	0.237	0.242	0.283	0.263	0.268	0.362	0.334	0.340
				SE.EST	0.247	0.234	0.237	0.285	0.268	0.271	0.353	0.331	0.334
				CP95	0.882	0.956	0.955	0.925	0.957	0.962	0.937	0.962	0.958

Table 3

Simulation results under $MAR_{TX,Q}$ for complete case analysis (CCA), our method (L^*_Q), multiple imputation (MI) and the estimating equations approach (GR). True values for the parameters were $\beta_1 = 0.223 = \log(1.25)$ and $\beta_2 = 0.916 = \log(2.5)$. Sample size for each simulation iteration was 1,000. We present relative bias (%), empirical standard deviation (SD.EMP), median estimated standard error (SE.EST) and confidence interval coverage rate with 95% confidence level (CP95).

$e\%$	%Missing	Subtype	β_1 (true value 0.223)				β_2 (true value 0.916)			
			CCA	L_Q^*	MI	GR	CCA	L_Q^*	MI	GR
1	49.5%	Bias(%)	109.870	-4.370	-5.460	-4.550	29.700	-0.870	-0.030	-5.530
		SD.EMP	0.268	0.234	0.241	0.242	0.224	0.172	0.177	0.185
		SE.EST	0.272	0.237	0.243	0.252	0.212	0.168	0.171	0.190
		CP95	0.855	0.953	0.959	0.945	0.761	0.937	0.937	0.951
1.25	47.2%	Bias(%)	101.010	-7.510	-9.400	-11.840	28.360	-1.040	-0.090	-5.880
		SD.EMP	0.288	0.240	0.247	0.252	0.208	0.170	0.172	0.186
		SE.EST	0.269	0.235	0.241	0.252	0.208	0.168	0.171	0.190
		CP95	0.864	0.951	0.950	0.939	0.794	0.950	0.954	0.963
1.75	44.1%	Bias(%)	94.690	-6.000	-6.250	-15.100	26.910	0.510	1.240	-5.080
		SD.EMP	0.260	0.225	0.232	0.234	0.208	0.169	0.172	0.191
		SE.EST	0.263	0.231	0.235	0.248	0.199	0.166	0.168	0.188
		CP95	0.884	0.963	0.962	0.958	0.767	0.948	0.947	0.955
2.5	41.0%	Bias(%)	92.680	1.690	1.220	-10.500	22.020	-0.920	-0.160	-6.940
		SD.EMP	0.271	0.231	0.239	0.245	0.205	0.171	0.175	0.190
		SE.EST	0.259	0.228	0.233	0.246	0.192	0.164	0.166	0.183
		CP95	0.874	0.950	0.947	0.944	0.809	0.940	0.937	0.940
5	36.1%	Bias(%)	78.480	-0.920	-1.630	-17.900	17.010	-0.840	-0.050	-8.670
		SD.EMP	0.265	0.235	0.241	0.248	0.179	0.158	0.161	0.176
		SE.EST	0.253	0.224	0.228	0.242	0.183	0.162	0.163	0.180
		CP95	0.890	0.947	0.948	0.946	0.871	0.956	0.955	0.957

Table 4

Simulation results under NMAR for complete case analysis (CCA), our method (L^*_Q), multiple imputation (MI) and the estimating equation approach (GR). True values for the parameters were $\beta_1 = 0.223 = \log(1.25)$ and $\beta_2 = 0.916 = \log(2.5)$. Sample size for each simulation iteration was 1,000. We present relative bias (%), empirical standard deviation (SD.EMP), median estimated standard error (SE.EST) and confidence interval coverage rate with 95% confidence level (CP95).

$e\%$	%Missing	Subtype	β_1 (true value 0.223)			β_2 (true value 0.916)				
			CCA	L_Q^*	MI	GR	CCA	L_Q^*	MI	GR
1.25	46.1%	Bias(%)	103.650	-5.040	-3.990	-8.850	26.840	-2.370	-1.840	-7.490
		SD.EMP	0.275	0.240	0.251	0.250	0.192	0.160	0.164	0.177
		SE.EST	0.272	0.241	0.245	0.255	0.202	0.163	0.165	0.186
		CP95	0.860	0.956	0.956	0.952	0.788	0.954	0.952	0.960
1.75	41.0%	Bias(%)	96.430	-6.950	-5.370	-15.930	19.890	-6.240	-5.750	-11.780
		SD.EMP	0.267	0.235	0.245	0.246	0.188	0.155	0.159	0.184
		SE.EST	0.273	0.245	0.250	0.262	0.188	0.156	0.157	0.181
		CP95	0.883	0.972	0.971	0.964	0.845	0.929	0.929	0.934
2.5	36.2%	Bias(%)	93.620	-0.600	1.660	-15.710	15.870	-7.690	-7.250	-13.690
		SD.EMP	0.272	0.244	0.249	0.254	0.177	0.149	0.151	0.178
		SE.EST	0.272	0.248	0.251	0.262	0.177	0.151	0.152	0.180
		CP95	0.872	0.961	0.962	0.954	0.877	0.929	0.933	0.934

Risk factor analysis of CRC according to MSI status using a complete case analysis (CCA) and our proposed method L_Q^* . The examined risk factors are family history of CRC (Family), BMI, regular aspirin use (Aspirin) and cumulative pack-year of smoking (SmokePkyr).

Table 5

CCA					
	$\hat{\beta}(\hat{\beta})$	$\widehat{SE}(\hat{\beta})$	p	$\hat{\beta}(\hat{\beta})$	L_Q^* $\widehat{SE}(\hat{\beta})$ p
BMI	0.013(1.013)	0.017	0.47	0.016(1.016)	0.014 0.27
Aspirin	-0.279(0.757)	0.196	0.15	-0.163(0.850)	0.145 0.26
MSI					
SmokePkyr	0.013(1.013)	0.003	< 0.001	0.014(1.014)	0.002 < 0.001
Family	0.738(2.091)	0.151	< 0.001	0.605(1.832)	0.134 < 0.001
BMI	0.026(1.026)	0.008	0.002	0.017(1.017)	0.006 0.003
Aspirin	-0.186(0.831)	0.100	0.063	-0.253(0.777)	0.063 < 0.001
MSS					
SmokePkyr	0.006(1.006)	0.002	0.002	0.007(1.007)	0.001 < 0.001
Family	0.274(1.316)	0.099	0.006	0.200(1.221)	0.072 0.006