Marginalized Frailty-Based Illness-Death Model: Application to the UK-Biobank Survival Data

Malka Gorfine a, Nir Keret a, Asaf Ben Arie a, David Zucker b, and Li Hsu c

 a Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel; b Statistics and Data Science, The Hebrew University, Jerusalem, Israel; c Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA

ABSTRACT
The UK Biobank is a large-scale health resource comprising genetic, environmental, and medical information on approximately 500,000 volunteer participants in the United Kingdom, recruited at ages 40–69 during the years 2006–2010. The project monitors the health and well-being of its participants. This work demonstrates how these data can be used to yield the building blocks for an interpretable risk-prediction model, in a semiparametric fashion, based on known genetic and environmental risk factors of various chronic diseases, such as colorectal cancer. An illness-death model is adopted, which inherently is a semi-competing risks model, since death can censor the disease, but not vice versa. Using a shared-frailty approach to account for the dependence between time to disease diagnosis and time to death, we provide a new illness-death model that assumes Cox models for the marginal hazard functions. The recruitment procedure used in this study introduces delayed entry to the data. An additional challenge arising from the recruitment procedure is that information coming from both prevalent and incident cases must be aggregated. Lastly, we do not observe any deaths prior to the minimal recruitment age, 40. In this work, we provide an estimation procedure for our new illness-death model that overcomes all the above challenges. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

1. Introduction
The UK Biobank (UKB) is a large-scale health resource comprising genetic, medical, and environmental information on approximately 500,000 volunteer participants in the United Kingdom, recruited at ages 40–69 during the years 2006–2010. The project monitors the health and well-being of its participants, thus providing a strong incentive for joining. The participants have undergone various measurements, provided blood, urine, and saliva samples for future analysis, and have also provided detailed information about themselves. In using these data to study a given disease, the participants can be classified into three groups: those already diagnosed with the disease at the time of recruitment ("prevalent" cases), those in whom the disease is diagnosed during follow-up ("incident" cases), and those not diagnosed with the disease as of the end of follow-up.

One concern in such a design is delayed entry, since subjects need to live at least up to the minimum recruitment age to participate in the study. Moreover, the presence of prevalent cases in the data requires special attention due to two reasons: (i) The delayed-entry correction for those observations must be different to that for the incident cases. (ii) The data are subject to recall bias—when participants are asked to provide information regarding their status at the time of disease diagnosis, it is likely that some information will be reported inaccurately, especially if a long time has passed since then. Death before the disease constitutes a competing risk to many studied phenotypes, and while death in general can censor the disease, the disease cannot censor death, hence these settings are termed "semi-competing risks."

Colorectal cancer (CRC) is the fourth most common type of cancer and the second most common cause of cancer-related deaths in the UK (Virdee et al. 2019). The purpose of this work is to establish the building blocks required for a reliable CRC risk prediction tool that accounts for individual susceptibility and has sound parameters’ interpretation, both for CRC diagnosis and for post-diagnosis death. The method accommodates environmental and genetic risk factors, death as a competing risk and data ascertainment. This is in contrast to existing CRC UKB data analyses (Morris et al. 2018; Bradbury, Murphy, and Key 2020) where the association between CRC and environmental factors is the main goal. We believe that the UKB data is a highly practical and important resource for producing risk-prediction models for rare complex diseases, such as CRC, as long as the ascertainment and competing risks are correctly accommodated. To our knowledge, we are the first to analyze the CRC UKB data to provide a risk-prediction model for CRC diagnosis and death after CRC diagnosis, which includes environmental and genetic risk factors.

Out of 484,918 UKB participants with available genetic and environmental data, there is a total number of 5131 CRC cases, of which 2339 are prevalent and 2792 are incident. The number of deaths before disease diagnosis is 12,767, and the number
of deaths after CRC diagnosis is 1040. As the study continues, the number of prevalent cases will no longer change because recruitment is already complete.

Under semi-competing risks settings, three stochastic processes are typically studied: the time until disease onset, time until death free of the disease, and time until death after disease onset (see Figure 1). Fine, Jiang, and Chappell (2001) considered a gamma frailty model defined on the upper wedge of the joint distribution of age at diagnosis and age at death, and supplied a consistent estimator for the parameter of the gamma distribution, but did not incorporate covariates. Xu, Kalbfleisch, and Tai (2010) proposed gamma-frailty illness-death conditional and marginal models in which the conditional (on the frailty variate) model is a Cox-type model. If the marginal distribution is of primary interest, interpretation of the regression effects in their model is cumbersome as the hazards do not take a simple form and also involve the frailty distribution parameter. Chen (2012) assumed a semi-parametric transformation model for the marginal regressions and a copula model for the joint distribution. However, it was assumed that occurrence of the nonterminal event does not alter the distribution of the terminal event, which is unrealistic in most illness-death scenarios. Extending the copula model to account for the change of distribution is not straightforward. Vakulenko-Lagun and Mandel (2016) considered an illness-death model with delayed entry and right-censored data, under a fully parametric regression framework, including a known functional form for the effect of age at diagnosis on death. Vakulenko-Lagun, Mandel, and Goldberg (2017) described a nonparametric inverse probability weighting (IPW) approach for estimating the joint distribution of disease and death times, subject to right censoring and delayed entry. This approach does not incorporate covariates. In addition, under the sampling scheme present in the UKB data, an IPW approach is inapplicable (as will be explained in Section 3.3). Zhou et al. (2016) described a simple pseudo-likelihood approach with copulas, aimed at estimating the marginal survival functions and the association parameter of the copula, but did not account for delayed entry and did not incorporate covariates. The approach of pseudo-values was presented by Andersen and Klein (2007), to directly estimate the covariate effects on the state probabilities, using a generalized estimating equations procedure. This approach requires the user to predetermine the time grid for the state probabilities. In addition, calculation of the pseudo-values for a dataset as big as the UKB poses a big computational burden.

In illness-death models, age at death after disease diagnosis is left truncated by the age at diagnosis. In most applications, it is unrealistic to assume that the observed covariates contain all the sources of dependence between age at diagnosis and age at death after diagnosis. Limited literature exists on Cox regression with dependent left-truncation and right-censored data. A complete review can be found in Shen (2017). In Section 4.2, we present an analysis of the UKB data with age at diagnosis included as a covariate in the model of age at death after diagnosis, in the spirit of Mackenzie (2012) and Shen (2017). The counterintuitive results of these analyses will be discussed (Section 4).

None of the aforementioned approaches provides a satisfactory framework for the analysis of the UKB data. We are seeking a model that can accommodate delayed entry and a dependence structure between the three aforementioned processes. In addition, we want a model that is easy to interpret at the population level and can incorporate risk factors and individual susceptibility as covariates. Lastly, we would advise against using a fully parametric model, as these models require more assumptions and are typically less robust.

The novelty of this work consists of several aspects: (i) Formulation of a new frailty-based illness-death model with Cox-type marginalized hazards. Under random sampling and right-censored data, the model parameters are consistently estimated. (ii) Adjusting the proposed estimators to accommodate delayed entry and the presence of both prevalent and incident cases, as in the UKB data. (iii) Analysis of the CRC UKB data yielding a risk-prediction model of CRC diagnosis and death after CRC diagnosis, based on environmental and genetic risk factors. R code for the data analysis and reported simulations is available at Github site: https://github.com/nirkeret/frailty-LTRC.

The rest of the article is organized as follows. Section 2 describes our proposed frailty-based illness-death model, and the pseudo-likelihood estimation approach for simple cohort studies with no delayed entry. The estimation procedure for delayed-entry data is outlined in Section 3. Section 4 is dedicated to the analysis of CRC UKB data. Section 5 summarizes simulation results. Final remarks are presented in Section 6.

2. Methods: Right-Censored Data, No Delayed Entry

2.1. Models

Two types of hazard functions are considered: a conditional hazard given the unobserved frailty variate and the observed time-independent covariates, and a marginalized hazard function with respect to the frailty variate, namely, the hazard given the time-independent covariates. The main focus is in estimating the illness-death marginalized hazards and survival functions. We adopt the approach of Glidden and Self (1999), and reformulate the frailty model so that the marginalized hazard functions obey a specified model, such as the Cox model, that is free of the frailty distribution parameter. In the context of this work, it is preferable that the marginal model be free of the frailty parameter, to facilitate its interpretation as a model.
corresponding to a randomly selected individual from the population. The frailty distribution parameter quantifies the degree of dependence between the different processes within the same person, so when it is present in a marginal model it obscures the interpretation of the regression coefficients.

The three processes, the time until disease onset, time until death free of the disease, and time until death after disease onset (see Figure 1), are sometimes measured on a sojourn time scale, namely, the disease-death process is not expressed as age of death after disease, but rather as the amount of time spent in diseased state until death. However, a negative association is expected between age at diagnosis and the amount of time spent in diseased state until death (residual life after diagnosis at older age tends to be shorter than that of diagnosis at younger age).

Since we wish to use a shared random effect to describe the dependence between the processes, it is natural to work with the age-scale and avoid situations with negative dependence.

Define T1 and T2 be age at diagnosis and age at death, respectively. Denote the unobserved frailty by a random variable \( \omega > 0 \) with a cumulative distribution function \( F \) indexed by an unknown parameter \( \theta \). Let \( Z \) be a vector of time-independent covariates. Based on the notation of Figure 1, let the conditional hazards of transition from state 1 to either state \( k = 2 \) or 3, given \((Z, \omega)\), be

\[
\begin{align*}
    h_{1k}^c(t|Z, \omega) &= \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(t \leq T_{k-1} < t + \Delta|T_1 \geq t, T_2 \geq t, Z, \omega) \\
    &= \omega \alpha_{1k}(t|Z) \quad t > 0 \quad k = 2, 3.
\end{align*}
\]

Let the conditional hazard function of leaving state 2, given \((Z, \omega)\), and given that the transition to state 2 occurred at age \( t_1 \) be

\[
\begin{align*}
    h_{23}^c(t|t_1, Z, \omega) &= \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(t \leq T_2 < t + \Delta|T_1 = t_1, T_2 \geq t, Z, \omega) \\
    &= \omega \alpha_{23}(t|Z) \quad t > t_1 > 0.
\end{align*}
\]

The nonnegative functions \( \alpha_{jk}, jk = 12, 13, 23 \), will be determined by the distribution of the frailty \( \omega \) and the marginalized hazards presented below. The frailty distribution \( F \) should be chosen such that the hazard models are identifiable. This is typically done via a parameterization that forces the expectation of the frailty variable to be equal to 1.

The corresponding Cox marginalized hazards are defined by

\[
\begin{align*}
    h_{1k}(t|Z) &= \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(t \leq T_{k-1} < t + \Delta|T_1 \geq t, T_2 \geq t, Z) \\
    &= h_{01k}(t) \exp(\gamma_{1k}^T Z) \quad t > 0 \quad k = 2, 3,
\end{align*}
\]

and

\[
\begin{align*}
    h_{23}(t|t_1, Z) &= \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(t \leq T_2 < t + \Delta|T_1 = t_1, T_2 \geq t, Z) \\
    &= h_{023}(t) \exp(\gamma_{23}^T Z) \quad t > t_1 > 0,
\end{align*}
\]

where \( \gamma_{jk} \) and \( h_{0jk}, jk = 12, 13, 23 \), are the regression coefficient vectors and unspecified baseline hazard functions, respectively. In \( h_{23} \), the disease onset time \( t_1 \) is not included in the vector of covariates, but instead the dependence between \( T_1 \) and \( T_2 \) comes from two sources: the so-called explanatory hazard ratio (Xu, Kalbfleisch, and Tai 2010), \( h_{023}/h_{013} \), and the frailty variate \( \omega \). If \( T_2 \) does not depend on \( T_1 \), one expects that the explanatory hazard ratio \( h_{023}(t)/h_{013}(t) = 1 \) for all \( t > 0 \), and that \( \omega \) is also 1. Our goal is estimating the regression coefficients \( \gamma_{jk} \), the hazards, \( h_{0jk}, jk = 12, 13, 23 \), and the dependence parameter \( \theta \).

In contrast to the above marginalized approach, Xu, Kalbfleisch, and Tai (2010) defined a frailty-based illness-death model in which the conditional hazards follow Cox-type models multiplied by a frailty variate, while the marginalized hazards are functions of the frailty-distribution parameter (Xu, Kalbfleisch, and Tai 2010, eqs. (19) and (20)).

If one is interested in estimating only \( \gamma_{jk} \) and \( h_{0jk} \) with \( jk = 12, 13 \), the standard partial likelihood approach can be applied, as in standard applications of Cox models with competing risks (Kalbfleisch and Prentice 2011, chap. 8). Estimation of \( \gamma_{23} \) and \( h_{023} \) could be more involved since age at death is left-truncated by the age at disease diagnosis. Under the assumption that \( T_1 \) and \( T_2 \) after diagnosis are conditionally quasi-independent given \( Z \) (Tsi 1990), \( \gamma_{23} \) and \( h_{023} \) can be easily estimated using a standard partial likelihood approach and a Breslow estimator, with the usual risk-set correction for left-truncated data [see Section S10 of the supplementary materials (SM)]. However, in most applications this assumption is untenable, because one cannot observe all the environmental and genetic effects that fully explain the dependence between \( T_1 \) and \( T_2 \). As a result, the standard approach yields biased estimators of \( \gamma_{23} \) and \( h_{023} \), as will be demonstrated in the simulation study (Section 4).

By this, we circumvent the dependent left-truncation problem. Note that \( \omega \) can also be interpreted as individual susceptibility underlying health conditions. We provide a unified estimation procedure for all the parameters of interest, including the dependence parameter.

As a final step of our new illness-death model presentation, we derive the relationship between \( \alpha_{jk} \) and \( h_{0jk}, jk = 12, 13, 23 \), for a given frailty distribution with differentiable inverse Laplace transform. Denote by \( \psi(s) = E(\exp(-s \omega)) \) the Laplace transform of \( \omega \), by \( \phi_k(s) \) its \( k \)th derivative with respect to \( s \), by \( \psi(s) \) the inverse Laplace transform, by \( \psi(k)(s) \) its \( k \)th derivative, by \( \xi(s) \) the inverse of \( -\phi(1) \), and finally by \( \xi(1)(s) \) the first derivative of \( \xi(s) \). Also, let \( h_{01k}(t) = \int_0^t h_{0jk}(u)du, H_{01k}(t|Z) = \int_0^t h_{1k}(u|Z)du \) and \( H_{12}(t|Z) = H_{12}(t|Z) = H_{12}(t|Z) \), \( jk = 12, 13, 23 \).

**Lemma 1.** For \( t > 0 \), the relationships between \( \alpha_{jk} \) and \( h_{0jk}, jk = 12, 13, 23 \), are given by

\[
\begin{align*}
    \alpha_{1k}(t|Z) &= -h_{01k}(t) \exp(\gamma_{1k}^T Z) \exp(-H_{11}(t|Z)), k = 2, 3,
\end{align*}
\]

and

\[
\begin{align*}
    \alpha_{23}(t|t_1, Z) &= -\xi(1) \exp(-H_{023}(t) \exp(\gamma_{23}^T Z)) \exp(-H_{023}(t) \exp(\gamma_{23}^T Z)) \exp(h_{023}(t)).
\end{align*}
\]

The proof of Lemma 1 is provided in Section S1 of the SM. Generally, a marginalized proportional-hazards model does not yield a conditional proportional hazards model. That is, \( \alpha_{jk} \) are of the form

\[
\alpha_{jk}(t|Z) = h_{0jk}(t)\alpha_{jk}^*(t|Z), \quad jk = 12, 13, 23,
\]
where $\alpha_{ik}^*\gamma$, which can be derived from the Laplace transform of the frailty distribution, depends on both $Z$ and $t$. For example, under the gamma-frailty model with expectation 1 and variance $\gamma$, $\phi(s) = (1 + \theta s)^{-1}/\gamma$, and thus $\alpha_{ik}^*\gamma(t|Z) = \exp(\gamma_{ik}^T Z) \exp(\theta H(t|Z))$, $k = 2, 3$, and $\alpha_{ik}^2(t|Z) = \exp(\gamma_{ik}^T Z) \exp[H_{23}(t|Z)\gamma(1 + \theta)/1 + \theta)]/(1 + \theta)$. As will be shown, the representation in Equation (1) plays an important role in our proposed estimation procedure.

### 2.2. Estimation Procedure

Suppose there are $n$ independent subjects. For the $i$th subject, $i = 1, \ldots, n$, denote by $C_i$ the censoring time. Let $V_i = (T_{i1} \wedge T_{i2}) \wedge C_i, \delta_{i1} = I(T_{i1} \leq (T_{i2} \wedge C_i))$, so that $\delta_{i1}$ equals 1 if the subject was observed to have the disease before being censored or dying. Also let $\delta_{i2} = I(T_{i2} \leq (T_{i1} \wedge C_i))$, so that $\delta_{i2}$ equals 1 if the subject died before having the disease or being censored. Denote by $W_i = \delta_{i1}(T_{i2} \wedge C_i)$ the age at death or age at censoring after having the disease, and $\delta_{i3} = \delta_{i1}I(T_{i2} \leq C_i)$, which equals 1 if death after the disease was observed. Then, the observed data consists of $(V_i, \delta_{i1}, \delta_{i2}, W_i, \delta_{i3}, Z_i), i = 1, \ldots, n$. The unobserved frailties $\omega_i, i = 1, \ldots, n$, are assumed to be independent random variables, independent of $Z_i$, with a cumulative distribution function $F$ indexed by an unknown parameter $\gamma$ and Laplace transform such that $\psi^{(1)}$ and $\ell^{(1)}$ exist. It is assumed that given $(Z_i, \omega_i)$, censoring and failure times are independent. Also, the censoring is noninformative for the frailty distribution and all other parameters in the models.

Let $\gamma = \gamma^{T}_{12}, \gamma^{T}_{12}, \gamma^{T}_{23}$ and $H_0 = (H_{012}, H_{013}, H_{023})$. The regression coefficients $\gamma$ will be estimated by maximizing a pseudo likelihood, while the cumulative hazard functions $H_0$ will be estimated with Breslow-type estimators. Since the likelihood contains $H_0$, and the Breslow-type estimator in turn requires $\gamma$ and $\theta$, a circular dependence is created, which calls for an iterative algorithm. The proposed estimation procedure is an extension of Gorfine, Zucker, and Hsu (2006) for the standard shared-frailty model for correlated failure times. A discussion that compares the proposed method and that of Gorfine, Zucker, and Hsu (2006) is provided in Section S2 of the SM.

The likelihood function is proportional to $L(\gamma, \theta, H_0) = \prod_{i=1}^n L_i$, where

$$
L_i = \int f(V_i, \delta_{i1}, \delta_{i2}, W_i, \delta_{i3}|Z_i, \omega) \exp(\gamma_{i1}^T Z_i) \exp(\theta_{i1}^T H(t|Z_i)) d\omega \\
\times \{h_{012}(V_i) \alpha_{i2}^*(V_i|Z_i) \exp(-1)^i \sum_{j=1}^{23} \exp(\gamma_{i2}^T Z_i) \exp(\theta_{i2}^T H(t|Z_i)) d\omega \\
\times \{h_{023}(W_i) \alpha_{i3}^*(W_i|Z_i) \exp(-1)^i \sum_{k=1}^{23} \exp(\gamma_{i3}^T Z_i) \exp(\theta_{i3}^T H(t|Z_i)) d\omega},
$$

$\delta_{i} = \sum_{j=1}^{23} \delta_{ij}$, $A_{ik}(t|Z) = \int_0^t \alpha_{ik}(u|Z) \exp(\gamma_{ik}^T Z) \exp(\theta_{ik}^T H(t|Z)) d\omega$, $i = 12, 13, 23$, and

$$
s_i = A_{12}(V_i|Z_i) + A_{13}(W_i|Z_i) + \delta_{i1} A_{23}(W_i|Z_i)
- \delta_{i1} A_{23}(Z_i|Z_i).
$$

A detailed explanation of Equation (2) is provided in Section S3 of the SM. Finally, for a given estimator of $H_0$, denoted by $\hat{H}_0$, the pseudo maximum likelihood estimator of $(\gamma, \theta)$ is defined to be the argument which maximizes $L(\gamma, \theta, \hat{H}_0)$.

Estimation of $H_0$ will be done by applying the estimation theorem (Aalen 1978, Theorem 3.4). We start with defining three counting processes. Let $\Xi$ be the maximal follow-up time, and for any $t \in [0, \Xi]$ and $i = 1, \ldots, n$ define the counting processes

$N_{i12}(t) = \delta_{i1}I(V_i \leq t)$, $N_{i13}(t) = \delta_{i2}I(V_i \leq t)$, and $N_{i23}(t) = \delta_{i3}I(W_i \leq t)$.

A key assumption is that given the covariates and the frailty variate, $N_{i12}$ and $N_{i13}$ are independent. Each $N_{ijk}(t)$ has conditional stochastic intensity process

$$
Y_{ijk}(t) = \alpha_{ijk}(t|Z_i) \omega_i = Y_{ij}(t) h_{ijk}(t) \alpha_{ijk}(t|Z_i) \omega_i
$$

where $Y_{ij}(t) = I(V_i \geq t)$ and $Y_{ij}(t) = \delta_{i1}I(V_i \leq t \leq W_i)$. The $\sigma$-algebra, denoted by $F_t^{(1)}$, generated by the observed history related to $j = 12, 13$ up to time $t$ is defined by

$$
F_t^{(1)} = \sigma\{N_{i12}(s), N_{i13}(s), Y_{ij}(s), Z_i, i = 1, \ldots, n; 0 \leq s \leq t\}.
$$

The $\sigma$-algebra related to $jk = 23$ consists of those observations that were diagnosed with the disease, namely, those with $\delta_{i1} = 1$ and $\delta_{i2} = 0$, so it is defined by

$$
F_t^{(2)} = \sigma\{N_{i23}(s), Y_{ij}(s), Z_i, i = 1, \ldots, n; 0 \leq s \leq t\}.
$$

By the innovation theorem, the unconditional stochastic intensity process of $N_{i1k}(t)$, $k = 2, 3$, with respect to $F_t^{(1)}$ is given by

$$
E(\omega_i | F_t^{(1)}) = \int_0^\Xi \omega dF(\omega | F_t^{(1)})
$$

$$
= \int \omega \exp[-\omega A_{i1}(t|Z_i)] d\omega = \int \omega \exp[-\omega A_{i1}(t|Z_i)] d\omega
$$

$$
= (-1)^{N_{i1}(t)+1} \Phi^{N_{i1}(t)+1} A_{i1}(t|Z_i)
$$

$$
= (-1)^{N_{i1}(t)+1} \Phi^{N_{i1}(t)+1} A_{i1}(t|Z_i),
$$

$N_{i1}(t) = N_{i12}(t) + N_{i13}(t)$ and $A_{i1}(t|Z_i) = A_{i12}(t \wedge V_i|Z_i) + A_{i13}(t \wedge V_i|Z_i)$. For subjects with $\delta_{i1} = 1$ and $\delta_{i2} = 0$, the unconditional stochastic intensity process of $N_{i23}(t)$ with respect to $F_t^{(2)}$ is given by

$$
E(\omega_i | F_t^{(2)}) = \int_0^\Xi \omega dF(\omega | F_t^{(2)})
$$

$$
= \int \omega \exp[-\omega A_{i2}(t|Z_i)] d\omega
$$

$$
= \int \omega \exp[-\omega A_{i2}(t|Z_i)] d\omega
$$

$$
= (-1)^{N_{i2}(t)+1} \Phi^{N_{i2}(t)+1} A_{i2}(t|Z_i)
$$

$$
= (-1)^{N_{i2}(t)+1} \Phi^{N_{i2}(t)+1} A_{i2}(t|Z_i),
$$

and $A_{i2}(t|Z_i) = A_{i12}(t \wedge V_i|Z_i) + A_{i23}(V_i \wedge t|Z_i) - A_{i23}(V_i|Z_i)$.

For example, under the gamma frailty, $E(\omega_i | F_t^{(1)}) = \theta^{-1} + N_{i1}(t)/\theta^{-1} + A_{i1}(t|Z_i)$, and when $\delta_{i1} = 1$ and $\delta_{i2} = 0$, $E(\omega_i | F_t^{(2)}) = \theta^{-1} + 1 + N_{i23}(t)/\theta^{-1} + A_{i2}(t|Z_i)$.
The respective Breslow-type estimators of $H_{ijk}(\cdot)$, $jk = 12, 13, 23$, are defined as step functions with jumps at the respective observed failure times. That is,

$$
\hat{H}_{ijk}(t) = \sum_{t_{i} \leq t} \Delta \hat{H}_{ijk}(s_{i}), \quad jk = 12, 13, 23,
$$

(4)

the $s_{i}$'s are the observed failure times, with

$$
\Delta \hat{H}_{01k}(t) = \frac{\sum_{i=1}^{n} \delta_{k-1} I(V_{i} = t)}{\sum_{i=1}^{n} Y_{i}(t) \hat{\alpha}_{k}^{*} (t - |Z_{i}| \hat{E} \left( \omega_{|}\hat{F}_{t|}^{(1)} \right)}, \quad k = 2, 3,
$$

(5)

and

$$
\Delta \hat{H}_{023}(t) = \frac{\sum_{i=1}^{n} \delta_{3} I(W_{i} = t)}{\sum_{i=1}^{n} Y_{i}(t) \hat{\alpha}_{23}^{*} (t - |Z_{i}| \hat{E} \left( \omega_{|}\hat{F}_{t|}^{(2)} \right)},
$$

(6)

where in $\hat{E}(\omega_{|}\hat{F}_{t|}^{(j)})$, $j = 1, 2$, and in $\hat{\alpha}_{jk}^{*}$, $jk = 12, 13, 23$, the unknown parameters are replaced by their estimators. A detailed description of $\hat{A}_{jk}$, the estimators of $A_{jk}$, $jk = 12, 13, 23$, is provided in Section S4 of the SM.

The proposed estimation procedure is summarized as follows.

Step 1. Use standard Cox regression software to obtain initial values of $\hat{\gamma}_{12}$, $\hat{\gamma}_{13}$ and $\hat{\gamma}_{23}$, by running three separate models, and take $\hat{\theta}$ to a value near independence.

Step 2. Use the current values of $(\hat{\gamma}_{T}, \hat{\theta})$ and estimate $H_{0ijk}$, $jk = 12, 13, 23$, by Equations (4)–(6).

Step 3. Use the current $\hat{H}_{0ijk}$, $jk = 12, 13, 23$, and estimate $(\gamma_{T}, \theta)$ by maximizing $L(\gamma_{T}, \theta, \hat{H}_{0})$.

Step 4. Iterate between Steps 2 and 3 until convergence is reached.

Let $\mu = (\gamma_{T}, \theta)^{T}$, $\hat{\mu} = (\hat{\gamma}_{T}, \hat{\theta})^{T}$, $\mu^{o} = (\gamma_{T}^{o}, \theta^{o})^{T}$, and $H_{0}^{o} = (H_{012}^{o}, H_{013}^{o}, H_{023}^{o})$, where the superscript $o$ denotes the respective true value. The following theorem summarizes the asymptotic properties of the proposed estimators. The required technical conditions and a sketch of the proof are provided in Section S5 of the SM.

**Theorem 1.** Under the assumptions listed in Section S5 of the SM, $\hat{\mu}$ is a consistent estimator of $\mu$, sup $|\hat{H}_{0ijk}(t) - H_{0ijk}(t)| = O_{p}(n^{-1/2})$, $jk = 12, 13, 23$, $\sqrt{n}(\hat{\mu} - \mu^{o})$ is asymptotically mean-zero multivariate normal, and $\sqrt{n}(\hat{H}_{0ijk}(t) - H_{0ijk}(t)), jk = 12, 13, 23$, converges to a Gaussian process.

### 2.3 Variance Estimation

Deriving the asymptotic or finite-sample variances of the proposed estimators analytically, is challenging, and is not attempted here. Instead, we advocate the use of the weighted bootstrap approach (Kosorok, Lee, and Fine 2004). Within each bootstrap sample, a random weight is assigned to each observation, from a standard exponential distribution. The estimators of each bootstrap sample are then derived based on $\log L(\gamma, \theta, H_{0}) = \sum_{i=1}^{n} \eta_{i}^{(b)} \log(L_{i})$, where $\eta_{i}^{(b)}$ is the weight for subject $i$ of the $b$th bootstrap repetition. Likewise, the $b$th bootstrap estimation of the baseline hazard function $H_{012}(t)$ consists of

$$
\Delta \hat{H}_{012}^{(b)}(t) = \frac{\sum_{i=1}^{n} \eta_{i}^{(b)} \delta_{1} I(V_{i} = t)}{\sum_{i=1}^{n} \eta_{i}^{(b)} Y_{i}(t) \hat{G}_{023}^{(b)}(t - |Z_{i}| \hat{E} (\omega_{|}\hat{F}_{t|}^{(1)}))},
$$

and similarly for $H_{013}(t)$ and $H_{023}(t)$. The weighted bootstrap approach is more suitable than regular bootstrap for survival data, because the regular bootstrap leads to tied event times and may, with heavily censored data, produce samples with a low number of events.

### 2.4 Computational Aspects

We analyze the large-scale UKB dataset. Taking CRC as an example, among the 221,723 men (263,195 women) there were 1603 (1189) CRC incident cases, 7752 (5015) died during the follow-up time before having CRC, and out of the 2945 (2186) prevalent and incident CRC observations, 668 (372) died. Thus, 212,368 men and 256,991 women were censored. Our estimation procedure with such a big sample size is time consuming. Accordingly, below we describe a simple technique for reducing the sample size with only a small efficiency loss, in the spirit of the basic ideas used in case-cohort designs (Cai and Zeng 2004). In particular, the log-likelihood based on (2) can be written as

$$
\sum_{i=1}^{n} \left[ \sum_{j=1}^{12} \delta_{ij} \log[h_{0j+1}(V_{i})\alpha_{j+1}^{*}(V_{i}|Z_{i})] \\
+ \delta_{3j} \log[h_{023}(W_{i})\alpha_{23}^{*}(W_{i}|Z_{i})] \right] \\
+ \sum_{i=1}^{n} I(\delta_{i} > 0) \log((-1)^{\delta_{i} \phi(\delta_{i})}(s_{i})) \\
+ \sum_{i=1}^{n} I(\delta_{i} = 0) \log \phi(s_{i}),
$$

(7)

where $s_{i}$ is given in (3). In the CRC UKB data, the last sum consists of more than 200,000 observations, within each sex, while most of the information is provided by the events. Let $n_{0} = \sum_{i=1}^{n} I(\delta_{i} = 0)$. Then, for big datasets with high censoring rates, we recommend taking a random subsample of size $\tilde{n}$ among the censored observations (i.e., those with $\delta_{i} = 0$), denoted by $\tilde{C}$, and the above log-likelihood function is replaced by

$$
\sum_{i=1}^{\tilde{n}} \left[ \sum_{j=1}^{12} \delta_{ij} \log[h_{0j+1}(V_{i})\alpha_{j+1}^{*}(V_{i}|Z_{i})] \\
+ \delta_{3j} \log[h_{023}(W_{i})\alpha_{23}^{*}(W_{i}|Z_{i})] \right] \\
+ \sum_{i=1}^{\tilde{n}} I(\delta_{i} > 0) \log((-1)^{\delta_{i} \phi(\delta_{i})}(s_{i})) \\
+ \frac{n_{0}}{\tilde{n}} \sum_{i=\tilde{n}+C}^{n} \log \phi(s_{i}).
$$
Similarly, the denominators of the cumulative baseline hazard estimators of \( H_{01k}, k = 2, 3 \), are replaced by

\[
\sum_{i=1}^{n} I(\delta_i > 0) Y_{i(t)}(t) \hat{\alpha}_{1k}^{*} (t - |Z_i| \tilde{E} (\alpha_i | \mathcal{F}_{-i}^{(1)}) \right. \\
+ n_0 \sum_{i \in C} Y_{i(t)}(t) \hat{\alpha}_{1k}^{*} (t - |Z_i| \tilde{E} (\alpha_i | \mathcal{F}_{-i}^{(1)}) .
\]

There is no change in the estimator of \( H_{023} \) since the subsampling step has no effect on the observations involved with this estimator. In summary, the sample consists of all the observations with at least one observed event and a random subsample from the censored data, where for each observation of the subsample a weight of \( n_0/n \) is assigned; the rest are assigned with a weight of 1. Section 5 includes some numerical results of efficiency loss as a function of \( \tilde{n} \). In our UKB data analysis, \( \tilde{n} = 20,000 \).

### 3. Methods: Right-Censored Data and Delayed Entry

#### 3.1. Data and Assumptions

Subject \( i \) is recruited at age \( R_i, c_L \leq R_i \leq c_U, i = 1, \ldots, n \), and then followed prospectively until death or censoring, whichever comes first. In the UKB data, \( c_L = 40 \) and \( c_U = 69 \). Thus, the data consist of \( n \) independent observations, each with \( (V_i, \delta_i, \delta_{2i}, W_i, \delta_{3i}, Z_i, R_i) \). Given \( (Z_i, \omega_i, R_i) \), it is assumed that the censoring is independent of the failure times and noninformative for the frailty distribution and all the other parameters in the model. Some participants had the disease before recruitment \( (R_i > V_i) \), and these observations are referred to as prevalent, whereas those who develop the disease after being recruited \( (R_i \leq V_i) \) are referred to as incident observations. Such a design, known also as length bias (or left truncation), suffers from sampling bias since only those individuals who live long enough are observed.

As there are no incident cases below the age of \( c_L \), one cannot estimate from the data any of the hazard functions below that age. In this work we focus on CRC. Since having CRC before age 40 is very rare (Ouakrim et al. 2015), we assume that the probability of having the disease before age \( c_L \) is practically zero. Hence, the estimators of \( h_{0jk}, jk = 12, 23 \), are only very slightly biased. Such an assumption should not be adopted for diseases such as breast cancer, where approximately 7% of diagnoses are before the age of 40 years (Anders et al. 2009). Thus, our proposed analysis is not directly applicable for such phenotypes and additional adjustment is required. Likewise, it is impossible to estimate the hazard function of death before having the disease directly from the observed data.

For an illness-death model with delayed entry, there are three principal statistical methods for inference (Vakulenko-Lagun and Mandel 2016, and references therein): (i) an unconditional approach where \( R_i \) is considered as a random variable with a known distribution, and its distribution is included in the likelihood function; (ii) a conditional approach where the value of the recruitment age is conditioned upon; or (iii) a conditional approach where the entire observed history up to the recruitment age is conditioned upon. In practice, multivariate survival data with delayed entry are most often analyzed using approach (iii) (Andersen 1988; Saarela, Kulathinal, and Karvanen 2009). In an illness-death model with approach (iii), prevalent individuals are not considered for estimation prior to their entry time, so they only contribute for estimating the parameters related to transition from diseased state to death.

Since we have no knowledge or reasonable assumptions on the recruitment distribution, approach (i) is inapplicable. While approach (ii) is more efficient than (iii) since it uses more information, it is more challenging computationally, as it requires an additional complicated numerical integration (see details in Section S6 of SM). In this section, we propose an estimation procedure adapted for delayed entry, which is based on the procedure of Section 2, and applies approach (iii).

Our proposed estimation procedure for accommodating delayed entry consists of three modifications: (1) adjusting the likelihood; (2) leveraging external information to estimate the baseline hazard function of disease-free death, \( h_{013} \), at age \( t < c_L \); and (3) adjusting the hazard function estimators, \( h_{0jk}, jk = 12, 13, 23 \).

#### 3.2. Adjusted Likelihood

The likelihood function based on the observed data given the history up to entry age \( R_i, i = 1, \ldots, n \), is given by \( L^T(Y, \theta, H_0) = \prod_{i \in R_i < V_i} L^T \prod_{i \in R_i \geq V_i} L^TT \), where

\[
L^T_i \propto \left( h_{012}(W_i) \alpha_{12}^{*} (V_i | Z_i) \right)^{s_{1i}} \left( h_{013}(W_i) \alpha_{13}^{*} (V_i | Z_i) \right)^{s_{3i}} \left( h_{023}(W_i) \alpha_{23}^{*} (W_i | Z_i) \right)^{s_{2i}} \left\{ 1 - \delta_i \phi(\epsilon^{(s_{1i})}, \phi(s_{2i})) \right\}^{s_{1i}} / \phi(s_{2i}),
\]

where \( s_{2i} = A_{12}(V_i | Z_i) + A_{13}(V_i | Z_i) + A_{23}(W_i | Z_i) - A_{12} (V_i | Z_i) \) and \( s_{2i} = A_{12}(V_i | Z_i) + A_{13}(V_i | Z_i) + A_{23}(W_i | Z_i) - A_{23} (V_i | Z_i) \). Details on the derivation of the above formulas are provided in Section S7 of the SM.

#### 3.3. Estimating \( h_{013}(t) \) for Age \( t < c_L \)

Estimation of the hazard functions under delayed entry is usually done in one of two approaches: risk-set correction or IPW. In the risk-set correction approach, the risk set at time \( t \) consists of those who recruited prior to time \( t \) and have not been censored or experienced any of the events by time \( t \), namely those with \( R_i \leq t \leq V_i \) (Klein and Moeschberger 2006, p. 313). Using this definition of risk-set, the Breslow-type estimators of \( h_{012} \) and \( h_{013} \) are now applicable. The prevalent cases do not contribute to estimation of the baseline hazards \( h_{012} \) and \( h_{013} \), but they can nevertheless contribute to the estimation of \( h_{023} \). By contrast, with the IPW approach, prevalent cases can contribute more in some settings (Chang and Tzeng 2006). However, one of the required assumptions for consistency of an IPW-based estimator, is the positivity assumption, namely, no subject can have probability zero for being recruited. In the UKB data, those who died prior to age 40 are with 0 probability of being recruited. Hence, the IPW approach cannot be adopted.
We propose to estimate $H_{013}$ for $t \leq c_L$ by leveraging external information on the death rate in the general population, for example from official life tables. We assume that the marginal death distribution in the general population approximates sufficiently the marginal death distribution among individuals free of the disease (a reasonable assumption for diseases that are rare among individuals of age $c_L$ or less), and that there exists a reasonable comparability between the general population and the UKB population (see Table S4 of the SM). The first step is to use general population data to estimate the marginal hazard $h_{13}(t)$ for $t \leq c_L$. For our analysis, we have used data published by the UK Office for National Statistics (https://www.ons.gov.uk).

Proceeding further, for $t \leq c_L$ the marginal death survival function in the general population can be expressed as

$$S_{13}(t) = \int \exp[-H_{013}(t-)] f(z) dz$$

and, differentiating with respect to $t$, the marginal density function is seen to be equal to

$$f_{13}(t) = S_{13}(t)h_{13}(t) = h_{013}(t)\int \exp(y_{13}^T z) [-H_{013}(t-)] f(z) dz.$$

Hence, the relationship between $h_{13}(t)$ and $h_{013}(t)$ is

$$h_{013}(t) = h_{13}(t)\frac{\int \exp[-H_{013}(t-)] f(z) dz}{\int \exp(y_{13}^T z) [-H_{013}(t-)] f(z) dz}.$$

Assuming that $Z$ in the cohort is representative of its distribution in the population, then, given $\hat{\gamma}$, an estimator for $h_{013}(t)$ can be defined as

$$\hat{h}_{013}(t) = h_{13}(t)\frac{\sum_{i=1}^n \exp[-\hat{H}_{013}(t-)] f(z_{zi})}{\sum_{i=1}^n \exp(\gamma_{13}^T z_i) \exp[-H_{013}(t-)] f(z_{zi})}.$$

(9)

where $\hat{H}_{013}(t)$ is estimated successively from 0 to $c_L$ at a prespecified equally-spaced grid of $\kappa$ points, $u_1, \ldots, u_\kappa$, and $H_{013}(t) = c_L/\kappa \sum_{u_i \leq t} \hat{H}_{013}(u_i)$. If recruitment starts at age $c_L$, the estimator of $H_{013}$ will be based on (9) up to age $c_L$, and then will continue with the following estimator provided in Section 3.4.

For diseases where the probability of onset before $c_L$ is not negligible, such as breast cancer, a similar approach can be implemented upon the availability of similar disease incidence information to estimate $h_{12}$ and $h_{12}'$ before $c_L$.

3.4. The Adjusted Hazard Function Estimators

For estimating the cumulative hazard functions, the intensity processes above are used with a correction to the risk sets. Specifically, the respective Breslow-type estimators of $H_{0jk}(\cdot)$, $jk = 12, 13, 23$, are

$$\Delta\tilde{H}_{012}(t) = \frac{\sum_{i=1}^n I(R_i < V_i) \delta_{11j}(V_i = t)}{\sum_{i=1}^n I(R_i \leq t \leq V_i) \delta_{11j}(t | Z_i) \hat{E}(\omega_i | F_{cL}^{(1)})}, \quad t > 0,$$

(10)

$$\Delta\tilde{H}_{013}(t) = \frac{\sum_{i=1}^n I(R_i < V_i) \delta_{11j}(V_i = t)}{\sum_{i=1}^n I(R_i \leq t \leq V_i) \delta_{11j}(t | Z_i) \hat{E}(\omega_i | F_{cL}^{(1)})}, \quad t \geq c_L,$$

(11)

and for $t > 0$,

$$\Delta\tilde{H}_{0123}(t) = \frac{\sum_{i=1}^n I(R_i < V_i) \delta_{1j}(V_i = t)}{\sum_{i=1}^n I(R_i \leq t \leq V_i) \delta_{1j}(t | Z_i) \hat{E}(\omega_i | F_{cL}^{(2)})} \hat{H}_{0jk}.$$

(12)

To summarize, the following are the updated Steps 2 and 3 of the proposed estimation procedure for delayed-entry and right-censored data (Steps 1 and 4 are the same as before):

Step 2. Use the current values of $(\gamma^T, \hat{\theta})$ and estimate $H_{0jk}, jk = 12, 13, 23$, by (9)–(12).

Step 3. Use the current estimate $\tilde{H}_{0jk}$, $jk = 12, 13, 23$, and estimate $\gamma$ and $\theta$ by maximizing $L^{LT}(\gamma, \theta, \tilde{H}_0)$.

For estimating the variance of the estimators, we suggest using the weighted bootstrap, as described in Section 2.3.

4. Analysis of UKB CRC Data

4.1. Data Processing

The failure times related to CRC were defined to be the age at first invasive CRC diagnosis, and death from CRC. CRC cases were identified according to the ICD10 codes (C180, C182–C189, C19, and C20) or the ICD9 codes (1530–1534, 1536–1541). Cancer of the appendix or noninvasive (in situ) CRC cases were excluded, as well as cases of carcinoid or related tumors (8240–8249) or lymphomas (9590–9729). The UKB is a cohort study, and right censoring could occur for a various reasons, such as emigration from the UK before death or CRC, participants demand to dropout before death or CRC (very rare), or participants being free of CRC or alive at the time of analysis (the majority). Therefore, the conditional independence and noninformative censoring assumptions are very likely to hold in the UKB data.

4.2. Analysis Results

Our goal in the CRC UKB analysis is to build a model for CRC diagnosis and death following CRC, while properly accommodating death as a semi-competing risk, using established genetic and baseline environmental risk factors for CRC. Since we are not interested in interpreting the effect size of individual risk factors (as these are well established), but the collective effect of risk factors on CRC risk and death, we followed the score-based approach of Jeon et al. (2018), summarizing these risk factors into two components: E-score for CRC associated lifestyle and environmental risk factors and G-score for known CRC associated single nucleotide polymorphisms.

Specifically, we fitted a Cox model with the age at diagnosis of CRC as the outcome and the recruitment age used for risk set correction. The following well-established CRC risk factors were included: sex, height, body mass index, education, smoking status, alcohol consumption, ibuprofen use, drug use, use of post-menopausal hormones (women only), and physical activity. Prevalent CRC cases were excluded. The results are provided in Table S1 of the SM. The E-score of each participant is defined as a linear combination of all risk factors, each weighted by its estimated regression coefficient. The E-scores
were subsequently standardized by a quantile transformation based on the E-score empirical cumulative distribution function of the CRC-free observations to improve its portability across studies/cohorts accounting for potential different data collection instruments used. The transformed E-scores were then entered into the illness-death model as a covariate. Table S1 provides the means and standard deviations of the transformed E-scores, by CRC status and sex.

Similarly, a genetic risk score (G-score) was derived based on 72 single-nucleotide polymorphisms (SNPs) that have been identified by GWAS to be associated with CRC (Jeon et al. 2018). Each SNP variable was coded as dosage based on the number of risk allele copies (0, 1, or 2) if the SNP is directly genotyped, and the expected number of copies if it is imputed. The G-score was developed in a similar manner to the E-score. A detailed description of the SNPs and the analysis results are provided in Tables S2 and S3 of the SM.

It is well known that women have a much lower risk of CRC. To allow for more precise estimates of the baseline hazard functions, we applied our proposed illness-death model separately to men and women. Section S8 and Table S4 of the SM compare the distribution in the UKB cohort with that in the general UK population for height, BMI, smoking status, and education, based on available information from the UK Office for National Statistics. The differences between the distributions are small, as required for using the UK Office for National Statistics data as reference death rates up to age 40 (Section 3.3).

The regression coefficient vectors γ12 and γ13 included both G-score and E-score, while γ23 included just the G-score. The E-score is not included in the transition model 2 → 3 since this part of the model was estimated using both prevalent and incident data, and environmental data on the prevalent cases were expected to be subject to substantial recall bias. We compared our method with the following Cox model analyses, in which the disease onset age was, in the spirit of Shen (2017), included in the model for the transition 2 → 3:

Cox I: Three separate Cox models were fitted. \( \{\gamma_{12}, H_{012}\} \) were estimated with CRC age at diagnosis as the outcome, \( R \) was used for risk set correction, and other events were treated as independent censoring. \( \{\gamma_{13}, H_{013}\} \) were estimated with age at death before CRC as the outcome, \( R \) was used for risk set correction, other events were treated as independent censoring. \( \{\gamma_{23}, H_{023}\} \) were estimated with age at death after CRC as the outcome, and age at CRC diagnosis and \( R \) were used for risk set correction. See Section S10 of the SM for the partial likelihoods.

Cox II: \( \{\gamma_{12}, H_{012}, \gamma_{13}, H_{013}\} \) were estimated as in Cox I. In \( \gamma_{23} \) the standardized age at diagnosis was added as a time-independent covariate, to deal with the fact that \( V \) is a dependent left-truncation time (Shen 2017).

Cox III: \( \{\gamma_{12}, H_{012}, \gamma_{13}, H_{013}\} \) were estimated as in Cox I. The effect of age of CRC diagnosis is included using a linear truncated spline with three knots, at the 25%, 50%, and 75% quantiles.

The results are presented in Table 1 and Figure 2. Clearly, the three Cox models provide identical results for transitions 1 → \( k, k = 2, 3 \). As expected, the proposed estimators of \( \{\gamma_{12}, H_{012}, \gamma_{13}\} \), are similar to the Cox model. Under Cox, the baseline hazard function \( H_{013} \) equals 0 for \( t \leq 40 \), so the Cox estimator of \( H_{013} \) is smaller than the proposed estimator. Since Cox II and III include age at diagnosis as a covariate, there are substantial differences among the estimators of \( H_{023} \), with extreme results under Cox III with the linear truncated spline. Under the proposed approach, the G-score and E-score coefficients are both significantly greater than zero for the health-to-diseased transition process. Additionally, the G-score for

| Table 1. UKB analysis results: regression coefficient estimates (standard error). |
|---------------------------------------------------------------|
| **Cox I** | **Cox II** | **Cox III** | **Proposed (20k censored)** |
| 221,723 men; 1603 CRC incident events; 7752 deaths before CRC; out of the 2945 with CRC (prevalent and incident) 668 died (out of them, 462 are incident cases) |
| \( \theta \) | G-score 12 | E-score 12 | G-score 13 | E-score 13 | G-score 23 | Spline Q1 | Spline Q2 | Spline Q3 |
| 1.358 (0.091) | 0.743 (0.089) | 0.051 (0.039) | 0.785 (0.041) | -0.003 (0.139) | 1.439 (0.097) | - | - | - |
| 0.072 (0.140) | 1.358 (0.091) | 0.743 (0.089) | 0.051 (0.039) | -0.072 (0.140) | -0.421 (0.131) | 0.421 (0.131) | - | - |
| 1.423 (0.219) | 1.423 (0.219) | 0.749 (0.050) | 0.421 (0.131) | -0.251 (0.414) | 0.429 (1.536) | - | - | - |
| 0.004 (0.517) | - | - | - | - | - | - | - | - |
| 2.297 (0.161,0.170) |
| 263,195 women; 1189 CRC incident events; 5015 deaths before CRC; out of the 2186 with CRC (prevalent and incident) 372 died (out of them, 291 are incident cases) |
| \( \theta \) | G-score 12 | E-score 12 | G-score 13 | E-score 13 | G-score 23 | Spline Q1 | Spline Q2 | Spline Q3 |
| 1.416 (0.106) | 0.260 (0.101) | -0.002 (0.049) | 0.650 (0.050) | -0.273 (0.184) | -2.065 (0.158) | - | - | - |
| 1.416 (0.106) | 0.260 (0.101) | -0.002 (0.049) | 0.650 (0.050) | -0.396 (0.184) | 2.424 (0.466) | - | - | - |
| 1.404 (0.143) | 0.260 (0.101) | -0.002 (0.049) | 0.650 (0.050) | -0.396 (0.184) | 2.424 (0.466) | - | - | - |
| 0.248 (0.105) | -0.002 (0.049) | 0.650 (0.050) | 0.632 (0.058) | -0.594 (0.701) | 0.001 (0.707) | - | - | - |
| 0.208 (0.163) | -0.273 (0.184) | -0.399 (0.184) | 0.208 (0.163) | 0.535 (0.682) | - | - | - | - |

NOTE: Significant covariates, based on a two-sided test at 0.05 significant level, are in bold.
CRC does not exhibit a significant effect on the healthy-to-dead transition process, but the corresponding E-score does. This result seems plausible because many CRC-related risk factors such as smoking status and alcohol consumption, are known to be related to death in general, and not only to CRC.

Under the three Cox models, the effect of the G-score of the diseased-to-dead process is null for men, and negative for women. That is, in the three Cox analyses for the women, the G-score is associated with a higher risk for CRC and a lower mortality risk after diagnosis. This result is counterintuitive. In contrast, our proposed analysis indicates that the G-score is also associated with a higher mortality risk after CRC diagnosis. The frailty parameter is approximately 2, for both men and women. The large standard error of $\hat{\theta}$ among men (0.480) was driven by few outliers in the bootstrap sample, while the median absolute deviation was 0.204. As a result we can deduce that there is a nonnegligible dependence between the processes which was not accounted for through the covariates. Under the three Cox models, the assumption of quasi-independent left-truncated time for death after CRC is most likely violated, despite including age at diagnosis as a covariate, and thus the Cox model analyses yield biased estimates.

Figure 3 presents the hazard ratio $\hat{h}_{23}(t)/\hat{h}_{13}(t)$ versus time $t$, stratified by sex, for three settings: baseline (Escore=Gscore=0), Escore=Gscore=0.25 and Escore=Gscore=0.75, for both Cox II and the proposed method. Under the proposed approach, it is evident that the explanatory hazard ratio is well above 1, indicating that time to death is highly correlated with time to
disease occurrence, and the earlier the disease onset the greater risk of dying earlier too, compared with non-diseased individuals. It is also of interest to note that $\theta \approx 2$, suggesting that allowing different hazard functions for death with and without prior disease is not enough to account for the dependence of $T_2$ on $T_1$. As a comparison, we also plotted the hazard ratio versus time for the Cox II model, where the age at diagnosis is included as a covariate to account for the dependence. While it shows a similar pattern for early age, the ratio for ages above 70 years old is below 1, suggesting a negative dependence, which is counterintuitive given what we know about CRC and death. Figure 3 demonstrates that Cox-II model provides unreasonable results not only under the baseline setting, but also under other E-score and G-score values.

Kendall’s $\tau$ (Kendall 1938) is a well-known global dependence measure and under the gamma frailty model, Kendall’s $\tau = \theta/\left(\theta + 2\right)$ (Hougaard 2000). In the CRC UKB analysis, the Kendall’s $\tau$ estimates for men and women are 0.495 (SE = 0.053) and 0.535 (SE = 0.017), respectively, indicating a strong positive dependence. An important local measure of dependence is the cross-ratio function (Clayton 1978), which allows investigation of how dependence changes over time. In our context of CRC diagnosis and death, it is given by $\zeta(t_1, t_2) = h(t_2|T_1 = t_1)/h(t_2|T_1 > t_1)$, where $h(t_2|T_1 = t_1)$ is the death hazard given CRC was diagnosed at age $t_1$, and $h(t_2|T_1 > t_1)$ is the death hazard given the subject is free of CRC by age $t_1$. When $\zeta(t_1, t_2) = 1$ for all $(t_1, t_2)$, the CRC age at diagnosis and age at death are independent. Under the gamma frailty model with mean 1 and variance $\theta$, $\zeta(t_1, t_2) = 1 + \theta$. For men $\hat{\zeta}(t_1, t_2) = 2.957$ and for women $\hat{\zeta}(t_1, t_2) = 3.297$. Hence, a subject diagnosed with CRC at age $t_1$ is expected to have about three times the risk of dying at subsequent times as compared to a patient free of CRC at age $t_1$. This indicates that CRC is strongly related to death.

5. Simulation Study

5.1. Data Generation

An extensive simulation study was performed to demonstrate the finite-sample properties of the proposed estimation procedures, with and without delayed entry. We assumed a gamma frailty model with expectation 1 and variance $\theta$, and a covariate vector $Z_T = (Z_1, Z_2, Z_3, Z_4)$, where each covariate was generated independently from Uniform(0,1). Given a sample of $n$ frailty values and covariate vectors, the failure times $T_1$ and $T_2$ were generated based on the conditional models $h_{jk}^{T_1}$, $jk = 12, 13, 23$ and Lemma 1 that provides $\alpha_{jk}$ in terms of the marginalized model. See Section S12.1 of the SM for a detailed
sampling description. The marginalized baseline hazard functions were set to be
\[
\begin{align*}
  h_{012}(t) &= 0.005I(0 \leq t < 0.05) + I(t \geq 0.05), \\
  h_{013}(t) &= 0.5I(0 \leq t < 0.05) + I(0.05 \leq t \leq 0.15) \\
  &+ 2I(t \geq 0.15),
\end{align*}
\]
and \( h_{023}(t) = I(t \geq 0.12) \). The regression coefficients were chosen to be \( \gamma_{21}^T = (2, 0.2, 0.05, 0), \gamma_{13}^T = (0.05, 1, 0, 0), \) and \( \gamma_{23}^T = (1, 0, 0, 0.5) \). Three levels of dependence were studied, \( \theta = 0, 1, \) or \( 2 \), corresponding to Kendall’s \( \tau \) values of \( 0, 0.33, \) and \( 0.5 \), respectively. Studying the independence case, \( \theta = 0 \), explores whether using our procedure results in a substantial efficiency loss compared to the standard delayed-entry adjusted partial likelihood.

Without delayed entry, all observations were followed since time \( 0 \). Under delayed entry settings, recruitment ages, \( R \), were randomly generated from Uniform\((c_L, c_U)\), where \( c_L = 0.05 \) and \( c_U = 0.15 \). For each sample, a large dataset consisting of \( \{T_1, T_2, R\} \) was first generated, and from those who were still alive at their recruitment age \( R \), \( n \) observations were randomly sampled. Next, for each observation an independent censoring time was generated from an exponential distribution with a rate parameter \( 2 \). In addition, an administrative censoring was imposed at time \( 0.61 \). For \( \theta = 1 \), about 25% of observations are censored before disease onset or death, and among those who were diseased, about 65% are censored before death. For \( \theta = 2 \), the corresponding numbers are about 27% and 75%.

### 5.2. Simulation Results

The following results are based on 100 repetitions for each configuration, and a sample size of 5000 individuals. The pseudo-likelihood function was maximized with the L-BFGS-B algorithm, as implemented in the optim function in R. Convergence of the algorithm was reached once the relative change in the pseudo-log-likelihood between two consecutive iterations went below \( 0.0001 \).

Under the cohort setting with no delayed entry (see Tables S5 and S6 of the SM), the proposed and the conventional Cox model yielded unbiased estimators for the parameters of \( h_{12} \) and \( h_{13} \); however, the conventional Cox model analysis yielded biased parameter estimates for \( h_{23} \), whereas the proposed estimator was unbiased. For example, with \( \theta = 2 \), the corresponding estimates of \( \gamma_{23,1} = 1 \) by Cox and the proposed approach are 0.628 (SE = 0.172) and 1.037 (SE = 0.181). Also, the true values of \( H_{023}(t) \) at \( t = 0.2, 0.4, 0.6 \), are 0.08, 0.28, and 0.48, while the corresponding estimates of Cox are 0.056, 0.164, and 0.257 (SEs are 0.010, 0.026, and 0.039). The respective numbers based on the proposed approach are 0.080, 0.280, and 0.484 (SEs are 0.013, 0.042, and 0.072). Evidently, the proposed approach performs well in terms of bias. Under \( \theta = 0 \) the efficiency loss is minimal, if any.

Tables S7 and S8 summarize the simulation results under delayed entry. We contrast between Cox with delayed-entry adjustment by the risk set approach (see Section S10 of the SM) and the proposed methods. Biased results are displayed in bold. Cox estimators of \( \gamma_{12} \) and \( \gamma_{13} \) are unbiased, as expected. Table S7 presents the biased Cox-model estimator of \( \gamma_{23,1} \) under \( \theta > 0 \). The estimator of \( \gamma_{23,4} \) is practically unbiased as its corresponding covariate is independent of all the other covariates in the model, while the corresponding covariate of \( \gamma_{23,1} \) also affects the disease age at onset. Table S8 presents the biased Cox-model results of \( H_{013} \) under any value of \( \theta \), due to also ignoring the fact that no death can be observed before \( c_L = 0.05 \). In contrast, our proposed estimators perform well in terms of bias and coverage rates. In addition, we observe that applying our method under the independence scenario does not result in any substantial efficiency loss. Table S9 of the SM shows simulation results under smaller sample sizes, \( n = 500 \) or 1000. Evidently, also under these sample sizes, our proposed approach performs reasonably well and much better than Cox with respect to the bias in estimating the disease-to-death process.

### 5.3. The Effect of Subsampling Censored Observations

For investigating the efficiency loss due to subsampling censored observations, we modified the censoring distribution to yield a higher censoring rate. Specifically, based on \( n = 25,000 \) observations and 100 replications, on average, 1313.53 (SD = 33.6) were diagnosed with the disease, and 1400.95 (SD = 33.6) and 520.72 (SD = 21.5) died before and after diagnosis, respectively. The mean number of censored observations \( (n_0) \) is 22,835.52. Tables S10–S13 of the SM show the results of the mean, empirical standard error, and relative efficiency of the regression coefficients and \( H_{013} \). For \( j_k = 12, 13, 23 \), under various values of \( \hat{n} \), the censored subsample size. The relative efficiency is defined as the ratio of the empirical variance based on the full data to the respective empirical variance based on the subsample. Evidently, the point estimates are very stable with \( \hat{n} \geq 8000 \). The relative efficiency related to \( \theta \) is above 0.9 even with \( \hat{n} = 4000 \), since the information related to \( \theta \) comes from observations with at least one observed event, and we subsample only censored observations. With \( \hat{n} = 16,000 \), the relative efficiency is close to or above 0.9 for all parameters. Clearly, the estimates related to transition \( 2 \rightarrow 3 \) have no efficiency loss under any value of \( \hat{n} \) since all observations that contribute to this transition have at least one observed event.

In the UKB data analysis, among men, there are 1603 with CRC, 7752 died before CRC and 668 died after CRC diagnosis. The analysis we provided is based on \( \hat{n} = 20,000 \). For women, there are 1189 with CRC, 5015 died before CRC and 372 died after CRC diagnosis. The analysis we provided is based on \( \hat{n} = 20,000 \). Therefore, since the disease of interest, CRC, is rare, the number of controls we chose should achieve over 90% efficiency. For death, the efficiency is even better than for the CRC, and the efficiency of \( \theta \) is not affected by the proposed subsampling procedure.

### 5.4. Frailty Distribution Misspecification—Sensitivity Analysis

The most popular frailty distributions are gamma, inverse Gaussian (IG), positive stable (PS), and log-normal. The log-normal distribution is very close to the IG distribution, and they express high dependence at intermediate ages (Hougaard 2000). High early dependence is expressed by the PS model and a high late dependence is by gamma. Hence, gamma and PS are two extreme ends of the dependence structure, whereas IG and
log-normal are inbetween. For sampling failure times based on our marginalized model, the derivative of the inverse of the first derivative of Laplace transform is required. The following sensitivity analysis does not include the log-normal distribution, as its Laplace transform has no closed form, which makes the sampling in our new model extremely complicated. Section S12 of the SM provides a detailed description of the sampling methods of gamma, IG and PS, under our marginalized model.

So far, frailty-based illness-death models have been studied in the literature only under gamma frailty with no sensitivity analysis, while frailty-based clustered survival models have been studied under other frailty distributions, including IG. Interestingly, in contrast to the clustered survival data, the IG distribution is not a good fit for the illness-death models as it is limited to Kendall’s $\tau < 0.5$. This point is demonstrated in Table S14 of the SM. This table summarizes the empirical Kendall’s $\tau$ when the data were generated based on the gamma or IG frailty, under two settings of $h_{023}$. The data were generated such that $T_1$ and $T_2$ have the specified value of Kendall’s $\tau$ (0.25 or 0.4). Now, we look at the empirical Kendall’s $\tau$ among the various pairs of times of the three processes, where $T_{12}$ is age at diagnosis, $T_{13}$ is age at death without the disease, and $T_{23}$ is age at death for those with the disease. Under both frailty distributions the empirical Kendall’s $\tau$ between $T_{12}$ and $T_{23}$ is higher, with the difference depending on the baseline hazard function of 2 → 3 and the left truncation of death age by diagnosis age. For example, under IG, Kendall’s $\tau = 0.25$ between $T_1$ and $T_2$ and high $h_{023}$, the empirical dependence between $T_{12}$ and $T_{23}$ is as high as 0.693, and therefore, an IG model that is limited to Kendall’s $\tau < 0.5$ cannot be fitted to these data. Frailties such as gamma and PS, with Kendall’s $\tau$ between 0 to 1, have no such problem. Hence, the following sensitivity analysis is based on the assumed gamma frailty while data are generated from the extreme setting of the PS distribution.

Table S15 of the SM shows that our proposed approach under gamma frailty is fairly robust even under the extreme frailty distribution PS. The empirical bias is usually not substantial, and the empirical coverage rates tend to be closer to the nominal level, compared to the Cox-based analysis, which corrects for left-truncation but ignores the dependence of age at death after diagnosis and age at diagnosis (i.e., dependent left truncation). For example, under PS with Kendall’s $\tau = 0.5$, $\gamma_{23,1} = 1$, the empirical means under Cox and our misspecified gamma frailty of $\gamma_{23,1}$ estimates are 0.176 and 0.723, respectively; the respective empirical coverage rates are 0.00 and 0.61. Moreover, when the baseline hazard function $h_{023}(t)$ increases from $I(t \geq 0.12)$ to $5I(t \geq 0.12)$, so that more deaths after diagnosis were observed, the empirical bias of our misspecified gamma frailty analysis reduced substantially, from 0.277 to 0.052. Interestingly, the simulation results reveal that although the frailty distribution is misspecified, Kendall’s tau estimator based on the gamma model, namely $\widehat{\theta}/(\widehat{\theta} + 2)$, has only a small empirical bias.

6. Discussion

We have proposed a novel semiparametric, shared-frailty based method for analyzing time-to-event data, within the illness-death model framework. Our model accounts for delayed entry and possible dependence between the stochastic processes, which are common in electronic medical records data such as the UKB data. The simulation study shows that the procedure works well in terms of bias and variance, and does not lose efficiency under independence. Importantly, from the UKB analysis, it can be seen that not properly accounting for possible dependence between the processes can lead to misleading results. Specifically, under the conventional Cox approaches, the genetic risk score is not associated with mortality risk after CRC diagnosis for men, and is associated with a lower mortality risk after diagnosis for women, whereas using our approach, the genetic risk score is associated with a higher mortality risk after CRC diagnosis.

The external data for estimating death rate under age 40 should be based on a representative data source and in some situations additional variation should be taken into account. In this article, the external death rate is based on the UK Office for National Statistics and as such, the variance of these incidence rates is expected to be very close to 0. Additionally, the UKB data are population-based and not research cohorts. Thus, selection bias, while expected for any cohort, is likely not a major issue.

An alternative estimation procedure, when there is no delayed entry, could be to estimate $\{\gamma_{12}, \gamma_{13}, H_{012}, H_{013}\}$ by a standard partial likelihood approach, estimating $H_{023}$ based on Equation (6), and then plugging these estimates in the likelihood, Equation (2), for estimating $(\theta, \gamma_{23})$, which is a different pseudo likelihood approach. An iterative procedure is still required, between the estimates of $(\theta, \gamma_{23})$ and that of $H_{023}$. Such a procedure might save some computation time but at the price of some efficiency loss. Under delayed entry, the estimator of $H_{013}$ would require a correction in the spirit of Section 3.2. In a future work, one can compare this approach with our proposed method. Establishing consistency and asymptotic normality would require new theoretical work.

Based on the proposed multistate model and estimators, predicting the absolute risk for developing CRC and for death after diagnosis can be developed based on

$$\Pr\left(T_1 \leq t^*, T_2 > T_1 | Z, T_1 > a, T_2 > a\right) = \frac{\int_{\max(0,a)}^{t^*} \int f_{12}(t|Z,\omega)S_{13}(t|Z,\omega)dF(\omega)dt}{\int S_{12}(a|Z,\omega)S_{13}(a|Z,\omega)dF(\omega)}$$

and

$$\Pr(T_2 \leq t^* | Z, T_1 = t, T_2 > a) = 1 - \frac{\int f_{12}(t|Z,\omega)S_{13}(t|Z,\omega)S_{23}(t^*|Z,\omega)dF(\omega)}{\int f_{12}(t|Z,\omega)S_{13}(t|Z,\omega)S_{23}(a|Z,\omega)dF(\omega)},$$

for a subject at a current age $a$ and any $t^* > a$. The predictions should be carefully evaluated by the standard criteria for a prediction model: calibration, accurate discrimination and sound clinical utility, while also considering frailty distribution misspecification.

Supplementary Materials

The Supplementary Materials include proof of Lemma 1 and main steps of the proof of Theorem 1; a comparison with Gorfine et al. (2009); details of the likelihood function, of $A_jk$ under gamma frailty model, and of Approach (ii); additional simulation results and additional results of the UKB data analysis.
**Funding**

The authors gratefully acknowledge support from the NIH (R01CA189532), the U.S.-Israel Binational Science Foundation (BSF, 2016126), and the Israel Science Foundation (ISF, 1067/17) in carrying out this work. This research has been conducted using the UK Biobank Resource, under project number 8614.

**References**

Aalen, O. (1988), “Nonparametric Inference for a Family of Counting Processes,” *The Annals of Statistics*, 6, 701–726. [1158]

Andersen, P. K. (1988), “Multistate Models in Survival Analysis: A Study of Nephropathy and Mortality in Diabetes,” *Statistics in Medicine*, 7, 661–670. [1160]

Andersen, P. K., and Klein, J. P. (2007), “Regression Analysis for Multistate Models Based on a Pseudo-Value Approach, With Applications to Bone Marrow Transplantation Studies,” *Scandinavian Journal of Statistics*, 34, 3–16. [1156]

Bradbury, K. E., Murphy, N., and Key, T. J. (2020), “Diet and Colorectal Cancer in UK Biobank: A Prospective Study,” *International Journal of Epidemiology*, 49, 246–258. [1155]

Cai, J., and Zeng, D. (2004), “Sample Size/Power Calculation for Case–Cohort Studies,” *Biometrics*, 60, 1015–1024. [1159]

Chang, S.-H., and Tzeng, S.-J. (2006), “Nonparametric Estimation of Sojourn Time Distributions for Truncated Serial Event Data—A Weight-Adjusted Approach,” *Lifetime Data Analysis*, 12, 53–67. [1160]

Chen, Y.-H. (2012), “Maximum Likelihood Analysis of Semicompeting Risks Data With Semiparametric Regression Models,” *Lifetime Data Analysis*, 18, 36–57. [1156]

Clayton, D. G. (1978), “A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence,” *Biometrika*, 65, 141–151. [1164]

Fine, J. P., Jiang, H., and Chappell, R. (2001), “On Semi-Competing Risks Data,” *Biometrika*, 88, 907–919. [1156]

Glidden, D. V., and Self, S. (1999), “Semiparametric Likelihood Estimation in the Clayton–Oakes Failure Time Model,” *Scandinavian Journal of Statistics*, 26, 363–372. [1156]

Gorfine, M., Zucker, D. M., and Hsu, L. (2006), “Prospective Survival Analysis With a General Semiparametric Shared Frailty Model: A Pseudo Full Likelihood Approach,” *Biometrika*, 93, 735–741. [1158]

Hougaard, P. (2000), *Analysis of Multivariate Survival Data*, New York: Springer-Verlag. [1164,1165]

Jean, J., Du, M., Schoen, R. E., Hoffmeister, M., Newcomb, P. A., Berndt, S. I., Caan, B., Campbell, P. T., Chan, A. T., Chang-Claude, J., and Giles, G. G. (2018), “Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors,” *Gastroenterology*, 154, 2152–2164. [1161,1162]

Kalbfleisch, J. D., and Prentice, R. L. (2011), *The Statistical Analysis of Failure Time Data* (2nd ed.), Hoboken, NJ: Wiley. [1157]

Kendall, M. G. (1938), “A New Measure of Rank Correlation,” *Biometrika*, 30, 81–93. [1164]

Klein, J. P., and Moeschberger, M. L. (2006), *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer. [1160]

Kosorok, M. R., Lee, B. L., and Fine, J. P. (2004), “Robust Inference for Univariate Proportional Hazards Frailty Regression Models,” *The Annals of Statistics*, 32, 1448–1491. [1159]

Mackenzie, T. (2012), “Survival Curve Estimation With Dependent Left Truncated Data Using Cox’s Model,” *The International Journal of Biostatistics*, 8, 29. [1156]

Morris, J. S., Bradbury, K. E., Cross, A. J., Gunter, M. J., and Murphy, N. (2018), “Physical Activity, Sedentary Behaviour and Colorectal Cancer Risk in the UK Biobank,” *British Journal of Cancer*, 118, 920. [1155]

Ouakrim, D. A., Pizot, C., Boniol, M., Malvezzi, M., Boniol, M., Negri, E., Bota, M., Jenkins, M. A., Bleiberg, H., and Autier, P. (2015), “Trends in Colorectal Cancer Mortality in Europe: Retrospective Analysis of the WHO Mortality Database,” BMJ, 351, h4970. [1160]

Saarela, O., Kulathinal, S., and Karvonen, J. (2009), “Joint Analysis of Prevalence and Incidence Data Using Conditional Likelihood,” *Biostatistics*, 10, 575–587. [1160]

Shen, P.-S. (2017), “Semiparametric Analysis of Transformation Models With Dependently Left-Truncated and Right-Censored Data,” *Communications in Statistics—Simulation and Computation*, 46, 2474–2487. [1156,1162]

Tsai, W.-Y. (1990), “Testing the Assumption of Independence of Truncation Time and Failure Time,” *Biometrika*, 77, 169–177. [1157]

Vakulenko-Lagun, B., and Mandel, M. (2016), “Comparing Estimation Approaches for the Illness-Death Model Under Left Truncation and Right Censoring,” *Statistics in Medicine*, 35, 1533–1548. [1156,1160]

Vakulenko-Lagun, B., Mandel, M., and Goldberg, Y. (2017), “Nonparametric Estimation in the Illness-Death Model Using Prevalent Data,” *Lifetime Data Analysis*, 23, 25–56. [1156]

Virdee, P. S., Kirtley, S., Elhussein, L., Watkinson, P. J., Holt, T. A., and Birks, J. (2019), “Components of the Full Blood Count as Risk Factors for Colorectal Cancer Detection: A Systematic Review Protocol,” *BMJ Open*, 9, e032759. [1155]

Xu, J., Kalbfleisch, J. D., and Tai, B. (2010), “Statistical Analysis of Illness-Death Processes and Semicompeting Risks Data,” *Biometrics*, 66, 716–725. [1156,1157]

Zhou, R., Zhu, H., Bondy, M., and Ning, J. (2016), “Semiparametric Model for Semi-Competing Risks Data With Application to Breast Cancer Study,” *Lifetime Data Analysis*, 22, 456–471. [1156]