Risk Projection for Time-to-Event Outcome Leveraging Summary Statistics With Source Individual-Level Data

Jiayin Zheng, Yingye Zheng, and Li Hsu

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA

ABSTRACT
Predicting risks of chronic diseases has become increasingly important in clinical practice. When a prediction model is developed in a cohort, there is a great interest to apply the model to other cohorts. Due to potential discrepancy in baseline disease incidences between different cohorts and shifts in patient composition, the risk predicted by the model built in the source cohort often under- or over-estimates the risk in a new cohort. In this article, we assume the relative risks of predictors are the same between the two cohorts, and propose a novel weighted estimating equation approach to recalibrating the projected risk for the targeted population through updating the baseline risk. The recalibration leverages the knowledge about survival probabilities for the disease of interest and competing events, and summary information of risk factors from the target population. We establish the consistency and asymptotic normality of the proposed estimators. Extensive simulation demonstrate that the proposed estimators are robust, even if the risk factor distributions differ between the source and target populations, and gain efficiency if they are the same, as long as the information from the target is precise. The method is illustrated with a recalibration of colorectal cancer prediction model.

1. Introduction
Accurate individual risk prediction for disease occurrence and progression is critical in clinical decision-making to identify high-risk patients that need medical intervention while preventing unnecessary treatment of low-risk patients (Alba et al. 2017). An important aspect of prediction model assessment is calibration, that is, the agreement between predicted risks and observed outcomes of the target population (Steyerberg et al. 2010). However, poor calibration occurs frequently in practice. For example, Collins and Altman (2012) found that the Framingham cardiovascular model over-estimated the risk in the United Kingdom (UK) patients; Abbasi et al. (2012) validated 12 prediction models for type 2 diabetes and most of them over-estimated the risk; and Drost et al. (2019) assessed two risk calculators for prostate cancer progression in five active surveillance cohorts, and found neither calculator provides consistent good risk estimates across five cohorts. If a model is used to inform patients, poorly calibrated risk estimates can lead to false expectations based on which patients may make personal decisions that were in fact misguided. Therefore, model recalibration is critically needed for accurate risk projection in the target population, but it is frequently under-appreciated (Van Calster and Vickers 2015).

Estimating disease risk involves two components: baseline risk and relative risks of risk factors. Suppose $T$ is time-to-event and $Z$ a $p \times 1$ vector of risk factors. Under the Cox proportional hazards model (Cox 1972), the probability of developing the disease before $t_1$ given being disease-free at $t_0$ in the absence of competing risks is

$$Pr(t_0 \leq T < t_1 | T \geq t_0, Z) = 1 - \exp\{\{\Lambda_0(t_0) - \Lambda_0(t_1)\} \exp(\beta_0^T Z)\},$$

where $\Lambda_0(t)$ is the cumulative baseline hazard function and $\exp(\beta_0)$ are the hazard ratios of $Z$. It is common to assume that $\beta_0$ is generalizable (Keiding and Louis 2016), however, $\Lambda_0(t)$ can be quite different between the source and the target cohorts, posing a major challenge when applying a model developed in one cohort to another.

In practice, sufficient individual-level information on both predictors and outcomes from the targeted cohort to allow a refit of the prediction model to obtain $\Lambda_1(t)$ is often not available. In contrast, information such as age-specific disease incidence rates and summary statistics (e.g., mean, quantile, and prevalence) of aspects of patient characteristics is often available from disease registries, historical patient data, and population-based survey data for the target population. Leveraging summary information on disease incidence rates has been proposed for obtaining the baseline hazard function (Gail et al. 1989; Liu et al. 2014). These approaches require a common joint distribution of all risk factors $f(Z)$ between the source cohort and the target population from which the disease incidence rates are obtained. However, this assumption is strong and often violated in practice. For example, in our real data application, many of the risk factors have very different frequencies across cohorts (Table 3 in...
Section 5). To relax this assumption, Chen et al. (2006) obtained $f(Z) = f(Z_1, Z_2)$ through combining the partial distribution $f(Z_1)$ from the target cohort and $f(Z_2|Z_1)$ from the source cohort, however, still assuming the conditional distribution is same in the two populations. This approach also requires individual-level data on $Z_1$ from the target population. Therefore, existing methods are limited in their assumptions on the covariate distributions and the requirement for individual level data from the target cohort when such assumptions are unjustified.

In this article, we propose a novel method to recalibrate the baseline hazard function for the target population, leveraging only the summary-level information without the need for a reference sample from the target population. Specifically, we propose an estimating equation for the target population (Qin and Lawless 1994; Owen 2001). While weights using the summary statistics of risk factors from the target population than the source risk factor distribution.

We employ the empirical likelihood (EL) approach to obtain the weights using the summary statistics of risk factors from the target population (Qin and Lawless 1994; Owen 2001). While EL has been considered in the literature, for example, to improve the efficiency of parameter estimation by considering external data sources (Chatterjee et al. 2016; Huang, Qin, and Tsai 2016), its adaptation to correct the bias due to difference in covariate distributions is novel and of practical interest especially in model recalibration. By solving the weighted estimating equation, we obtain an estimator that reduces the bias considerably when the risk factor distributions differ and gains efficiency if the summary information is precise.

In the rest of the article, we describe the proposed estimation procedure, large sample properties, and practical considerations in Section 2. We justify the EL weights in Section 3 and evaluate the finite sample performance through simulation in Section 4. An application to recalibrate a risk prediction model for colorectal cancer (CRC) developed in the Women’s Health Initiative (WHI) study (The Women’s Health Initiative Study Group 1998) to cohorts from the Prostate, Lung, Colorectal and Ovarian (PLCO) Trial study (Prorok et al. 2000) and the UK Biobank (UKB) (Sudlow et al. 2015) is presented in Section 5, and we conclude in Section 6. Technical details and additional numerical results are provided in the web-based supplementary materials.

2. Methods

2.1. Notation and Models

We denote the source population for this study by $P^s$, and refer $P^s$ and $E^s$ generically as the probability and expectation with respect to $P^s$. The interest is to generalize this model to a target population $P$, with the corresponding probability and expectation denoted by $P$ and $E$. We assume that individual-level data are only available for $P^s$, and for $P$ only some summary information about overall disease incidence rates and some risk factors. For both populations, we postulate the Cox model for the relationship between the failure time $T$ and risk factors $Z$.

For the source population $P^s$,

$$\lambda^s(t|Z) = \lambda_0(t) \exp(\beta^t Z),$$

where $\lambda^s(t|Z) = \lim_{\Delta t \to 0} \Pr(t < T < t + \Delta t | T \geq t, Z)/\Delta t$, $\lambda_0(t)$ is an unspecified baseline hazard function for $P^s$, and $\exp(\beta^t)$ is a $p \times 1$ vector of hazard ratio parameters. Following the convention, we assume the hazard ratios for $Z$ are common between the two populations. Then for the target population $P$, we have

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta^t Z),$$

where $\lambda(t|Z) = \lim_{\Delta t \to 0} \Pr(t < T < t + \Delta t | T \geq t, Z)/\Delta t$ and $\lambda_0(t)$ is an unspecified baseline hazard function for $P$. Note that $f(Z)$ is not necessarily same for $P$ and $P^s$.

For the source study let $T_i, L_i$, and $C_i$ be the failure time, the left truncation time (e.g., study entry), and the censoring time, respectively. Let $Z_i$ be a $p \times 1$ vector of risk factors in the prediction model. Define $X_i = \min(T_i, C_i)$ if $X_i \geq L_i$, and the disease indicator $I_i = I(L_i < T_i \leq C_i)$, where $I(\cdot)$ is an indicator function. We assume that $\{X_i, Z_i, i = 1, \ldots, n\}$ are independently and identically distributed (iid), and both $L_i$ and $C_i$ are independent of $T_i$ conditional on $Z_i$. Define the counting process $N_i(t) = I(L_i < X_i \leq t, \delta_i = 1)$ and the at-risk process $Y_i(t) = I(L_i < t \leq X_i)$. In addition, define $\tilde{N}(t) = n^{-1} \sum_{i=1}^n N_i(t)$, $H(t) = n^{-1} \sum_{i=1}^n Y_i(t)$, and $H_r(t; \beta) = n^{-1} \sum_{i=1}^n Y_i(t)Z^H_{i} \exp(\beta Z_i)$, where $r = 0, 1, 2$, $a^{0,1} = 1$, $a^{0,2} = a$, and $a^{1,2} = aa^T$. The maximum partial likelihood estimator $\hat{\beta}$ can be obtained by solving the score equations

$$U(\beta) = \sum_{i=1}^n \int_0^t \{Z_i - \tilde{Z}(t; \beta)\} dN_i(t),$$

where $\tilde{Z}(t; \beta) = H_1(t; \beta)/H_0(t; \beta)$ and $\tau$ is the end of the follow-up period such that $Pr(H(t) > 0) > 0$. The cumulative baseline hazard function $\tilde{A}_0(t) = \int_0^t \tilde{L}_0(s)\, ds$ can be estimated by the Breslow estimator (Breslow 1972), denoted as $\tilde{A}_0\hat{(}t\tilde{)} = \int_0^t \tilde{dN}(u)/H_0(u; \hat{\beta})$.

2.2. Proposed Estimation Methods

For recalibrating $A_0(t) = \int_0^t \tilde{A}_0(u) \, du$, the information about the disease-free probability $S(t) = Pr(T > t)$ from the target population $P$ is critical. Let $V(t) \in [0, \infty]$ be a generic cumulative baseline hazard function and $\Phi(Z; V(t), \beta, S(t)) = \exp(-V(t) \exp(\beta^t Z_i)) - S(t)$, for $t \in [0, \tau]$. It is easy to see that for target population, $E(\Phi(Z; A_0(t), \beta, S(t))) = 0$. Assuming $f(Z)$ is identical for $P$ and $P^s$, we estimate $E$ with an empirical estimator placing 1/n mass on each of the observed data points $\{Z_i, i = 1, \ldots, n\}$ from $P^s$. Suppose $S(t)$ is available at $t = t_1, \ldots, t_k$ from $P$, we recalibrate $A_0(t)$ at these time points by solving the estimating equation for $V(t)$:

$$n^{-1} \sum_{i=1}^n \Phi(Z_i; V(t), \hat{\beta}, S(t)) = 0 \quad (4)$$

$$n^{-1} \sum_{i=1}^n \left[\exp(-V(t) \exp(\hat{\beta}^t Z_i)) - S(t)\right] = 0.$$

Since $\Phi(Z; V(t), \beta, S(t))$ is a continuous and strictly decreasing function of $V(t)$ for any fixed $(Z, \beta, S(t))$, a unique nonnegative
solution to Equation (4) exists when \( S(t) > 0 \). We denote this solution as \( \hat{\Lambda}_0(t) \), which is called an unweighted estimator. The Newton-Raphson algorithm can be used to solve Equation (4). Note that the estimating equation \( E(\Phi(Z; \Lambda_0(t), \beta_0, S(t))) = 0 \) is equivalent to the integration of \( \hat{\lambda}(t) = (1 - AR(t)) / \lambda(t) \), where \( AR(t) \) is an attributable hazard function, which can be estimated from the case-control or cohort data (Liu et al. 2014; Pfeiffer and Gail 2017; Zhao et al. 2020), and \( \lambda(t) \) is the incidence rate for the target population.

The assumption of having a same risk factor distribution for both the source and the target populations may be stringent, implausible, and difficult to verify (Keiding and Louis 2016). If it is violated, \( \{Z_i, i = 1, \ldots, n \} \) from \( \mathcal{P}^* \) will not represent \( f(Z) \) in the target population \( \mathcal{P} \), and Equation (4) yields biased \( \hat{\Lambda}_0(t) \). To reduce the bias, a natural approach is to reassign the weights to \( \{Z_i, i = 1, \ldots, n \} \) in Equation (4), such that the weighted data are more representative of \( f(Z) \) in \( \mathcal{P} \). To achieve this, we will make use of available summary covariate information \( \mu_0 \) such that \( E(h(Z_i, \mu_0)) = 0 \), where \( h(Z, \mu) \equiv (h_1(Z, \mu), \ldots, h_q(Z, \mu))^T \) is a \( q \times 1 \) (known) mapping function with parameters \( \mu \). Common examples include \( h_1(Z, \mu_1) = Z_i - \mu_1 \) and \( h_2(Z, \mu_2) = (Z_i \leq \mu_2) - q \) where \( \mu_1 \) and \( \mu_2 \) are the mean and \( q \)th quantile for \( Z_i \), respectively. Using summary information \( \mu_0 \), we obtain weights by maximizing the log empirical likelihood (Qin and Lawless 1994; Owen 2001),

\[
\sum_{i=1}^n \log(w_i), \text{ under constraints}
\]

\[
\sum_{i=1}^n w_i h(Z_i, \mu_0) = 0, \quad w_i > 0, \quad \sum_{i=1}^n w_i = 1.
\]

Standard empirical likelihood programs can be used (see, e.g., http://statweb.stanford.edu/~owen/empirical/). As a common premise to empirical likelihood methods, the zero point vector \( \mathbf{0} \) should fall within the convex hull of \( \{h(Z_i, \mu_0), i = 1, \ldots, n\} \), we can now estimate \( \hat{\Lambda}_0(t) \) by solving the weighted estimating equation for \( V(t) \), \( n^{-1} \sum_{i=1}^n w_i \exp\{-V(t) \exp(\hat{\beta}^T Z_i) - S(t)\} = 0 \), for which a unique nonnegative solution \( \hat{\Lambda}^m_0(t) \) exists when \( S(t) > 0 \). Using the Lagrange multiplier, the weights satisfy \( \bar{w}_i = n^{-1}\{1 + \hat{\gamma}^T h(Z_i, \mu_0)^{-1}\} \), where \( \hat{\gamma} \), the Lagrange multiplier for the constraint \( \sum_{i=1}^n w_i h(Z_i, \mu_0) = 0 \), is the solution to \( \sum_{i=1}^n h(Z_i, \mu_0)/(1 + \hat{\gamma}^T h(Z_i, \mu_0)) = 0 \). This implies that \( \{\hat{\Lambda}^m_0(t), \hat{\gamma}\} \) are the solution to the equation \( \sum_{i=1}^n \rho(Z_i; V(t), \gamma, \mu_0, S(t), \hat{\beta}) = 0 \),

\[
\rho(Z, V, \gamma, \mu, S, \beta) = \frac{\exp\{-V(t) \exp(\hat{\beta}^T Z) - S(t)\}}{1 + \hat{\gamma}^T h(Z, \mu)} - \frac{h(Z, \mu)}{1 + \hat{\gamma}^T h(Z, \mu)}.
\]

Let \( \gamma_0 \) be the probability limits of \( \hat{\gamma} \) (its existence is proved in Web Appendix D). Essentially we have three risk factor distributions: \( f^\ast(Z) \) for \( \mathcal{P}^* \) from which the individual-level data are sampled, \( f(Z) \) for the target population \( \mathcal{P} \), and \( f^\ast(Z)/(1 + \gamma^T h(Z, \mu_0)) \) for an artificial population from which \( Z_i \)'s were sampled with weights \( 1/(1 + \gamma^T h(Z_i, \mu_0)) \), but \( f(Z) \) can be interpreted as the distribution close to \( f^\ast(Z) \), in an empirical likelihood sense, but subject to the restriction that \( \mu_0 \) is common with the target population. As shown later in Section 2.3, the estimator \( \hat{\Lambda}^m_0(t) \) obtained from solving (6) converges to \( \Lambda^0(t) \), which is the solution to \( E^\dagger\{\Phi(Z; V(t), \beta_0, S(t))\} = 0 \) where \( E^\dagger \) is the expectation taken over \( f^\ast(Z) \). With additional

Remark. The proposed recalibration approach does not require summary information from all the covariates. For computational simplicity one may filter out highly correlated constraints to avoid collinearity, and select key covariates to impose constraints on, in particular covariates that show large differences in summary statistics between populations and have strong effects. Real data analysis in Section 5 shows that incorporating the information of just a few key risk factors yields comparable performance to incorporating the information of all risk factors. Our extensive simulation also shows that simple summary statistics like mean, variance, median, or other quantiles are generally enough to yield almost unbiased estimator for the baseline hazard function. If the correlation structures of covariates are substantially different between the source and target populations, additional conditional/subgroup summary statistics may be needed to achieve good recalibration.

2.3. Large Sample Results

Let \( n \) be the sample size of the source study with individual-level data and \( m \) be the sample size of the target population sample that generates the summary statistics, and \( r = \lim_{n \to \infty} n/m \). The regularity conditions in the following theorems are listed in Web Appendix A. Below in Theorem 1 we first establish the consistency and asymptotic normality of \( \hat{\Lambda}^m_0(t) \) and \( \hat{\Lambda}^m_0(t) \) assuming the risk factor distribution is common between \( \mathcal{P} \) and \( \mathcal{P}^* \). We then relax this assumption in Theorem 2.

Theorem 1. Assume \( f(Z) = f^\ast(Z) \). Under conditions A1–A3, \( \hat{\Lambda}^m_0(t) \), \( \hat{\Lambda}^m_0(t) \) is consistent to \( \Lambda_0(t) \), where \( \Lambda_0(t) \) is the true cumulative baseline hazard function in the target population, and \( \sqrt{n} \{\hat{\Lambda}^m_0(t) - \Lambda_0(t)\} \overset{d}{\to} N(0, \sigma^2_0(t)) \). With additional conditions A5 and A6, we have \( \gamma_0 \to 0 \) and \( \hat{\Lambda}^m_0(t) \) is consistent to \( \Lambda_0(t) \). Further with condition A6, \( \sqrt{n} \{\hat{\Lambda}^m_0(t) - \Lambda_0(t)\} \overset{d}{\to} N(0, \sigma^2_0(t)) \). When \( n/m \) is low, that is, the sample size that generates the summary statistics is much larger than the source cohort, we show that \( \sigma^2_0(t) < \sigma^2_0(t) \), indicating that \( \hat{\Lambda}^m_0(t) \) is asymptotically more efficient than \( \hat{\Lambda}^0_0(t) \) (see Web Appendix C). Intuitively, the additional information imposes constraints on the full parameter space to a reduced parameter space, and if this additional information is precise, it will gain efficiency and the extent will depend on the precision of the additional information.

Theorem 2. Under conditions A1–A5, \( \hat{\Lambda}^m_0(t) \) is consistent to \( \Lambda_0(t) \), where \( \Lambda_0(t) \) is the solution to \( E^\dagger\{\Phi(Z; V(t), \beta_0, S(t))\} = 0 \) where \( E^\dagger \) is the expectation taken over \( f^\ast(Z) \). With additional
condition A6, $\sqrt{n}(\widehat{\Lambda}_0^w(t) - \Lambda_0^1(t)) \xrightarrow{d} N(0, \sigma_2^2(t))$ (see Web Appendix D).

Based on the asymptotic properties in Theorems 1 and 2, two $1 - \alpha$ confidence intervals (CI) of $\Lambda_0(t)$ can be constructed based on $\widehat{\sigma}_1(t)$ and $\widehat{\sigma}_2(t)$, which are the corresponding estimators of $\sigma_1(t)$ and $\sigma_2(t)$ obtained by replacing the expectations with empirical counterparts and parameters with their estimates.

In practice, the true values of $S(t)$ and $\mu_0$ in Equation (6) are usually replaced by their corresponding estimates $\hat{S}(t)$ and $\hat{\mu}$.

Let $\Sigma$ be the limiting covariance matrix of $\sqrt{m}\hat{\mu}$, and $\Sigma_{\mu,S(t)}$ the limiting covariance vector of $\sqrt{m}\hat{\mu}$ and $m\hat{S}(t)$. Variance $\sigma_2^2(t)$ includes $\Sigma$ and $\Sigma_{\mu,S(t)}$ as components. In practice, their estimates are often unavailable except for some diagonal elements of $\Sigma_{\mu}$ (i.e., the variance components). One may set all unavailable elements of $\Sigma_{\mu}$ and $\Sigma_{\mu,S(t)}$ to be 0. The resulting estimator, denoted as $\hat{\sigma}_2^2$, would approximate $\sigma_2^2$ well, if any one of the following three conditions holds: (1) all unavailable components are close to zero; (2) the ratio of sample sizes, $n/m$, is low; (3) the target sample size $m$ is not too small. Although the above asymptotic results were derived for a single value of $t$, it can be straightforwardly extended to allow for multiple values of $t$ based on the proof in Web Appendix D. In addition, the proposed inference procedure can be easily modified to accommodate the scenario where the summary information $(\hat{S}(t), \hat{\mu})$ are obtained from multiple sources.

### 2.4. Interpolation

We had focused on estimating $\Lambda_0(t)$ at the time points (say, $K$ points $t_1 < \ldots < t_K$) for which the disease probabilities from the target population are available. Using Equation (6) we obtain cumulative baseline hazard estimates at these time points, denoted by $\hat{\Lambda}_0(t_k), k = 1, \ldots, K$. If the interest is $\Lambda_0(t)$ where $t$ is in between two ages (say $t_{k-1}$ and $t_k$), one can interpolate $\Lambda_0(t)$ based on the estimated $\Lambda_0(t_{k-1})$ and $\Lambda_0(t_k)$. Note that $\hat{\Lambda}_0(t)$ is nondecreasing with $t$ and whether the interpolation between $t_{k-1}$ and $t_k$ is linear, convex or concave should be determined based on domain knowledge of the disease. For example, evidence based on SEER incidence rates supports using a convex interpolation for CRC probability. If there is no external information for the shape of interpolation, a natural approach to determining the shape is based on the baseline hazard from the source cohort. Accordingly, we propose an interpolation approach that postulates a specific relationship between the source $\hat{\Lambda}_0(t)$ and target $\Lambda_0(t)$ by $\Lambda_0(t) = g(\hat{\Lambda}_0(t_k), t; \phi)$ where $g(\cdot)$ is a continuous mapping function and $\phi$ is a vector of parameters. For example, assuming $g(\hat{\Lambda}_0(t_k), t; \phi) = \phi_0 \hat{\Lambda}_0(t_k)$ leads to a globally constant ratio relationship $\Lambda_0(t) = \phi_0 \Lambda_0(t_k)$. The ratio $\hat{\Lambda}_0(t_k)/\Lambda_0(t_k)$ can also be time-dependent, for example, the exponential trend $g(\hat{\Lambda}_0(t_k), t; \phi) = \exp(\phi_0 + \phi_1 t)$ or piecewise constant trend $g(\hat{\Lambda}_0(t_k), t; \phi) = \{\Lambda_0(t_k) - \Lambda_0(t_k-1)\} \exp(\phi_k) + \sum_{k=1}^{k-1} \{\Lambda_0(t_k) - \Lambda_0(t_{k-1})\} \exp(\phi_k) | t_{k-1} \leq t < t_K$ for $k = 1, \ldots, K + 1$ with $\tau_0 = 0$ and $\tau_{K+1} = +\infty$ where $\phi = (\phi_1, \ldots, \phi_{K+1})^T$. For piece-wise constant trend, $\phi_{K+1}$ is not estimable for $t > t_K$ and we assume $\phi_{K+1} = 0$.

### 2.5. Absolute Risk Estimation

Once we obtain $\hat{\Lambda}_0(t)$ and $\hat{\mu}$, we can calculate the absolute risk for developing disease for an individual given his/her risk profile. Sometimes the competing risks need to be taken into account. In our real data example of colorectal cancer, as the cancer tends to occur in older ages, individuals may die of other causes before developing colorectal cancer. As such, it is important to account for death when calculating the risk for developing colorectal cancer. For a person who is free of both events of interest and competing risks at age $t_0$ with a risk profile of $Z$, the absolute risk of experiencing the event of interest before age $t_1$ in the presence of competing event is defined as (Pfeiffer and Gail 2017)

$$Pr(t_0 < T \leq t_1, \Delta = 1 | T > t_0, Z) = \int_{t_0}^{t_1} \lambda(u|Z) \exp[- \int_{t_0}^{u} \{\lambda(s|Z) + \lambda_c(s|Z)\} ds] du,$$

where $T$ is the time to the first event, $\Delta$ is the event type with values of 1 and 2 indicating the event of interest and competing event, respectively, $\lambda(t|Z)$ and $\lambda_c(t|Z)$ are the respective cause-specific hazard functions at age $t$ given $Z$.

We can use the same method in Section 2 to estimate $\Lambda_0(t)$ for the event of interest in the presence of competing risks, even
when $Z$ is associated with the competing risks. In Equation (4), the summary statistic $S(t)$ from the target population may be replaced by the commonly used Kaplan–Meier estimator or exponent of the negative Nelson–Aalen estimator, obtained by treating individuals who experience competing events first as censored at that time. The theoretical results in Section 2.3, which assumes $Z$ having no effects on the competing risks, provide an approximated solution to calculating the variance of the baseline hazard estimator in the presence of competing risks. Subsequently, one can obtain the variance of the absolute risk estimator by the delta method. Extensive simulation results show that this approximation works well under a wide range of scenarios (Web Appendix J).

3. Justification for Empirical Likelihood Weights

We provide justification theoretically and numerically for the weighting scheme of empirical likelihood (EL) that the weighted distribution generally gets closer to the target than the unweighted one. The proofs are provided in Web Appendix I.

Assume $P^*$ and $P$ are the source and target distributions of $p$-dimensional random variable (covariate) vector $Z$ and both distributions have the same support $\mathcal{Z}$. Given observations $Z_i, i = 1, \ldots, n$ randomly sampled from $P^*$, maximizing the EL with constraints of subset statistics from $P$ yields an updated distribution $P^*_n(Z \leq c) = \sum_{i=1}^n w_i 1(Z_i \leq c)$ where $c = (c_1, \ldots, c_p)^T \in \mathcal{Z}$, $Z \leq c$ is element-wise inequalities, and $w_i$ are weights obtained by maximizing $\sum_{i=1}^n \log(w_i)$ under the constraints (5). Let $P^*(Z \leq c) = \lim_{n \to \infty} P^*_n(Z \leq c)$.

We study how close the EL-updated distribution $P^*$ to the target $P$ compared to the source $P^*$ by using the Kullback–Leibler (K-L) divergence (Kullback and Leibler 1951). The K-L divergence for $P$ and $P^*$ is defined as $D_{KL}(P||P^*) = \int f(z) \log(f(z)/f^*(z))dz$ where $f^*(z)$ and $f(z)$ are the probability densities/masses of $P^*$ and $P$ at $z$. The K-L divergence is always nonnegative and the closer is the K-L divergence to zero, the closer are the two distributions. We have the following general result.

**Theorem 3.** Assume the probability distributions of $Z_1$ and $(Z_1, Z_2)$ are available from the target population $P$, where $Z_1$ and $Z_2$ are exclusive subsets of variables from $Z$. Then the EL- updated distribution $P^*$ with the constraint $\mu_1 = f(Z_1)$ and $P^*$ with the constraint $\mu_2 = f(Z_1, Z_2)$ satisfy the following inequalities:

(a) $D_{KL}(P||P^*) \leq D_{KL}(P||P^*)$, where the equality holds if and only if $f(Z_1) = f^*(Z_1)$;

(b) $D_{KL}(P^*||P^*) \leq D_{KL}(P||P^*)$, where the equality holds if and only if $f(Z_2|Z_1) = f^*(Z_2|Z_1)$.

This theorem implies that constraining the distribution of any variables will yield an EL-updated distribution at least as close to the target distribution compared to the source distribution. If $P^*$ and $P$ have the same conditional distribution, adding constraints on $Z_2$ will not yield a closer distribution beyond constraints on $Z_1$. Below we provide more concrete examples for discrete and continuous variables, respectively.

3.1. Discrete Variables

Assume $Z$ is a vector of discrete random variables and denote $Z = (Z_1^T, Z_2^T)^T$. From Theorem 3, if the constraint of $f(Z_1)$ is imposed, then the EL-updated distribution $P^*$ is closer to $P$ than the source $P^*$. For example, if $Z_1$ is a binary variable, then imposing the prevalence $f(Z_1 = 1)$ will lead to $P^*$ closer to $P$ than $P^*$ in terms of K-L divergence, with the inequality strictly holds if $f^*(Z_1 = 1) \neq f(Z_1 = 1)$.

A common and practical scenario for a vector of $Z$ is that we have constraints on marginal probabilities from the target population but not the joint probabilities. Under this scenario, we show through extensive simulation that as long as the correlation structure of $Z$ in $P^*$ is not too far away from that in $P$, then $P^*$ is as close to and often closer to $P$ compared to the source $P^*$. For simplicity, assume $Z = (Z_1, Z_2)^T$, where $Z_1$ and $Z_2$ are binary variables. Let $p_j$ and $q_j$ denote the probability mass of $(Z_1, Z_2) = (i, j)$, $i, j = 0, 1$ for the source and target populations, respectively. Denote $p_i = p_{i0} + p_{i1}, q_i = q_{i0} + q_{i1}, p_j = p_{j0} + p_{j1}$ and $q_j = q_{j0} + q_{j1}$. We consider all possible combinations of $p_j$ and $q_j \in [0.05, 0.1, \ldots, 0.85]$. The reason for the upper bound of 0.85 is that each probability mass is at least 0.05 and thus the upper bound is $1 - 0.05 \times 3$. In total, we have $969 \times 969 = 938,961$ scenarios. We consider four distributions: the source $P^*$, the target $P$, the updated $P^*$ using $q_j$, and updated $P^*$ using $(q_1, q_1), P^*$ and $P$ have the same distribution of $Z_1$, and $P^*$ and $P$ have the same marginal distributions of $(Z_1, Z_2)$. We investigate the K-L divergence between $P$ and $P^*$, $(P^*, P^*)$, and $(P^*, P^*)$ in (a)–(c), as well as the differences $D_{KL}(P||P^*) - D_{KL}(P||P^*)$, $D_{KL}(P||P^*) - D_{KL}(P||P^*)$, and $D_{KL}(P||P^*) - D_{KL}(P||P^*)$ in (d)–(g) of all $938,961$ scenarios. By and large, $P^*$ is the closest to the target $P$, followed by $P^*$, and $P^*$ diverges most from $P$. As justified by Theorem 3, $D_{KL}(P||P^*) - D_{KL}(P||P^*) \geq 0$ for all the scenarios. Subfigure (e) shows $D_{KL}(P||P^*) - D_{KL}(P||P^*) \geq 0$ for all the scenarios except for very few where the difference is negligibly close to 0. Subfigure (f) shows that for most scenarios, $P^*$ using constraints of both marginal probabilities is closer to the target $P$, compared to $P^*$ that uses only one marginal probability as constraint. When we restrict the ratio of $OR_{P^*}$ and $OR_P$ to be within $[1/5, 5]$, where $OR_{P^*}$ and $OR_P$ are odds ratios between $Z_1$ and $Z_2$ for $P^*$ and $P$, respectively, the difference $D_{KL}(P||P^*) - D_{KL}(P||P^*) \geq 0$, indicating additional constraints, even if it is only marginal distribution, improves the updated distribution as long as the correlation among covariates is not extremely different between $P$ and $P^*$.

3.2. Continuous Variables

If $Z$ is a categorical variable, the probability masses at the distinct values are sufficient to specify the distribution of $Z$. For continuous $Z$, unless a parametric distribution is assumed, usually moments (e.g., mean and variance) or quantiles (e.g., median, Q1, Q3) are not sufficient to specify the entire distribution. As a result, it is difficult to have general asymptotic results. However, we are able to establish that if the quantile of $Z$ from $P$ is imposed for updating $P^*$, the EL-updated distribution
\[ D_{KL}(P \| P^*) \leq D_{KL}(P \| P^†) \]. In addition, we show through extensive simulation that under a wide range of scenarios imposing constraints of summary information (quantiles, moments or both) from \( P \), the updated \( P^† \) is as close or closer to \( P \) in terms of K-L divergence, compared to \( P^* \). Without loss of generality, we consider a univariate continuous variable \( Z \). We have the following results.

**Theorem 4.** Let \( \mu_j \) denote the \( q_j \)th quantile of \( P \), that is, \( P(Z \leq \mu_j) = q_j \) for \( j = 1, 2 \), and let \( P^† \) and \( P^‡ \) be the corresponding EL-updated distributions imposed with \( \mu_1 \) and \{\( \mu_1, \mu_2 \}\), respectively. The following holds:

(a) \( D_{KL}(P \| P^†) \leq D_{KL}(P \| P^*) \), where the equality holds if and only if \( P^*(Z \leq \mu_1) = q_1 \);

(b) \( D_{KL}(P \| P^‡) \leq D_{KL}(P \| P^†) \), where the equality holds if and only if \( P^*(Z \leq \mu_2) = q_2 \).

The proof can be obtained by noting that the p.d.f. of \( P^† \) is

\[ f^*(z) \frac{q_1}{p_1} \] for \( z \leq \mu_1 \) and \( f^*(z) \frac{(1 - q_1)}{(1 - p_1)} \) for \( z > \mu_1 \), where \( p_1 \equiv P^*(Z \leq \mu_1) \). The theorem implies that additional quantile information makes the updated distribution closer to the target, compared to the source.

Numerically, we considered eight common types of distribution with finite support, for both source and target distributions. They are representatively selected in terms of mean, variance, skewness, and kurtosis. We constructed these eight distributions by a mixture of beta and uniform distributions as shown in Figure 2 (top for source \( P^* \) and right for target \( P \)). We considered six EL-updated distributions with different constraints: \( P_{m1} \) (mean), \( P_{m2} \) (mean and variance), \( P_{q1} \) (median), \( P_{q2} \) (median, Q1 and Q3), \( P_{c1} \) (mean and median), \( P_{c2} \) (mean, variance, median, Q1, and Q3). We calculated seven K-L divergences of the source and six updated distributions to the target \( P \) for each combination of the source and target distributions, as shown for each cell in Figure 2. Among these constraints, \( D_{KL}(P \| P_{c2}) \) is often the lowest, indicating the EL-updated \( P_{c2} \) that imposes the constraints of both moments (mean, variance) and quantiles (median, Q1, Q3) is the closest to the target \( P \), while \( D_{KL}(P \| P^*) \) for the source distribution is often the largest. For all the scenarios, \( D_{KL}(P \| P_{m1}) \geq D_{KL}(P \| P_{m2}) \), \( D_{KL}(P \| P_{q1}) \geq D_{KL}(P \| P_{q2}) \) and \( D_{KL}(P \| P_{c1}) \geq D_{KL}(P \| P_{c2}) \), where \( \geq \) indicates greater than or similar to. These observations imply that imposing additional constraints on the variability (variance, Q1 and Q3) often makes the EL-updated distribution closer to the target distribution. For mean and median, usually having both constraints improves the K-L divergence than having only one with some exceptions, for example, when the source distribution is type 3 and the target distribution is type 2, 7, and 8 where the target is skewed or multi-modal and mean does not reflect well the density of distribution. Regarding the comparison between moment and quantile information, we have the following observations: when the source distribution is relatively close to uniform distribution (like type 1, 5, 6, and 8), constraints with mean is comparable with median but adding variance appears to be better than
Q1 and Q3; when the source distribution has relatively large kurtosis (like type 2, 3, 4, and 7), the quantile information are often better than moment. In practice, constraints with all five summary statistics (mean, variance, median, Q1, Q3) are often adequate even when the source distribution is substantially different from the target, based on our simulation.

4. Simulation Studies

We conducted extensive simulation to evaluate the finite sample performance of the proposed method under various scenarios. Specifically, we assumed that the effects of risk factors on the performance of the proposed method under various scenarios.

Specifically, we assumed that the effects of risk factors on the performance of the proposed method under various scenarios. We conducted extensive simulation to evaluate the finite sample summary statistics (mean, variance, median, Q1, Q3) are often better than moment. In practice, constraints with all five kurtosis (like type 2, 3, 4, and 7), the quantile information are often better than moment. In practice, constraints with all five summary statistics (mean, variance, median, Q1, Q3) are often adequate even when the source distribution is substantially different from the target, based on our simulation.

We generated two exposure variables \( Z = (Z_1, Z_2)^T \). For \( \mathcal{P}^* \), \( Z_1 \sim \text{Bernoulli}(0.5), Z_2 \sim N(0, 1) \) and \( Z_1 \perp Z_2 \). For \( \mathcal{P} \), we considered the following four configurations: (C1) \( f(Z) = f^*(Z) \); (C2) \( Z_1 \perp Z_2, f(Z_2) \) is same, but \( Z_1 \sim \text{Bernoulli}(0.8) \); (C3) \( f(Z_1) \) is same but \( Z_2|Z_1=1 \sim N(0.5, 1.2) \) and \( Z_2|Z_1=0 \sim N(-0.5, 0.8) \). We set the corresponding coefficients \( \beta_0 = \{ \log(2), \log(2) \} \). We generated right censoring time \( C = C^* I(1 \leq C^* \leq 100) + I(C^* < 1) + 100 I(C^* > 100) \), where \( C^* \sim N(40 + \zeta Z_1, 15), \zeta = -5, 0, 5 \). This yields approximately 64% to 85% censoring rates under various scenarios, representing moderate to high levels of censoring. The observed failure time was rounded to the nearest integer to mimic the real-world data. The sample size of the source study was \( n = 1000 \), representing moderate source cohort size. To generate the summary information from the target population, we generated a cohort from the target population and obtained the Kaplan–Meier estimates of disease-free probabilities and first- and second-moments of \( T(Z_1, Z_2) \). Note that the Kaplan–Meier estimator is not consistent given that censoring time is only conditionally independent of failure time given \( Z_1 \); however, this is a commonly available statistic for disease-free probabilities in practice and hence we used it here. For the sample size of the target cohort, we considered \( m = 200, 1000, \) and 100,000, representing small, moderate and large sample sizes of the external information resources. We generated 2,000 replicates for each simulation setting.

We compared six estimators: the Breslow estimator \( \hat{\Lambda}_B^w(t) \) based on the source study, the unweighted estimator \( \hat{\Lambda}_B^w(t) \), the weighted estimators \( \hat{\Lambda}_B^w(t) \) with four different constraints on \( Z \) each having more constraints than the previous one: \( \{E(Z_1), E(Z_2), E(Z_1), E(Z_2)\} \), and \( \{E(Z_1), E(Z_2), E(Z_1), E(Z_2)\} \), respectively. All methods except \( \hat{\Lambda}_B^w(t) \) incorporate disease-free probability estimates \( \hat{S}(t) \) at ages \( t = 1, 2, \ldots, 60 \) from the target population. The Breslow estimator \( \hat{\Lambda}_B^w(t) \) is served as basis for comparing the performance with other estimators, as it represents a conventional solution when no auxiliary information from the target population is used. We assessed their performance by the following metrics: percentage of...
Table 1. Summary statistics of simulation results for Scenario A1: $\Lambda_0(t) = \theta_0(t) = (0.01t)^2, n = 1000$, and $m = 100,000.$

| $t$ | $\Lambda_0(t)$ | $\tilde{A}_0(t)$ | $\tilde{A}_0^D(t)$ | $\tilde{A}_0^{W_0}(t)$ | $\tilde{A}_0^{W_3}(t)$ | $\tilde{A}_0^{W_4}(t)$ | $\hat{A}_0(t)$ | $\hat{A}_0^D(t)$ | $\hat{A}_0^{W_0}(t)$ | $\hat{A}_0^{W_3}(t)$ | $\hat{A}_0^{W_4}(t)$ |
|-----|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|----------------|----------------|----------------|----------------|----------------|
| 20  | 0.04           | 0.4%            | 1.1%            | 1.1%            | 1.1%            | 1.0%            | 0.2%           | 21.7%          | 2.5%           | 1.8%           | 1.8%           |
|     | ESD            | 0.65            | 0.44            | 0.44            | 0.43            | 0.42            | 0.65           | 0.52           | 0.58           | 0.54           | 0.54           |
|     | ASE            | 0.65            | 0.47            | 0.46            | 0.43            | 0.42            | 0.65           | 0.57           | 0.59           | 0.55           | 0.54           |
|     | sMSE           | 0.65            | 0.44            | 0.44            | 0.43            | 0.43            | 0.65           | 1.05           | 0.58           | 0.55           | 0.54           |
|     | sMSE           | 94.3%           | 96.3%           | 95.9%           | 95.6%           | 95.1%           | 94.6%          | 67.4%          | 95.9%          | 96.1%          | 95.7%          |
| 40  | 0.16           | 0.4%            | 1.0%            | 1.0%            | 1.0%            | 1.0%            | 0.2%           | 22.8%          | 2.4%           | 1.7%           | 1.7%           |
|     | ESD            | 1.89            | 1.53            | 1.52            | 1.47            | 1.47            | 1.90           | 1.77           | 2.12           | 1.98           | 1.98           |
|     | ASE            | 1.87            | 1.64            | 1.56            | 1.47            | 1.47            | 1.87           | 1.94           | 2.12           | 2.00           | 1.98           |
|     | sMSE           | 93.7%           | 96.2%           | 95.8%           | 95.3%           | 95.0%           | 93.1%          | 52.1%          | 95.2%          | 95.7%          | 95.7%          |
|     | CP             | 94.1%           | 90.8%           | 90.3%           | 90.3%           | 90.2%           | 91.3%          | 31.8%          | 44.6%          | 94.4%          | 95.2%          |
|     | ESD            | 1.90            | 1.54            | 1.53            | 1.48            | 1.48            | 1.90           | 4.14           | 2.16           | 1.99           | 2.00           |
|     | CP             | 91.8%           | 96.6%           | 94.8%           | 96.1%           | 95.7%           | 95.6%          | 95.5%          | 96.3%          | 95.7%          | 95.7%          |

NOTE: PBias, the relative bias; ESD, the empirical standard deviation ($\times100$ of the 2000 estimates; ASE, the mean of the asymptotic-based standard error ($\times100$); sMSE, the square root of mean squared error ($\times100$); CP, the coverage probability of a 95% confidence interval for $\Lambda_0(t)$; CAD, the average cumulative absolute deviation between the estimated and true cumulative hazard function across a time range of one through sixty.

relative bias (PBias), empirical standard deviation (ESD), asymptotic-based standard error (ASE), square root of mean squared error (sMSE), and coverage probability of 95% confidence intervals over the true value of $\Lambda_0(t)$ (CP) at selected ages $t = 20, 40, 60$. We also calculated the average cumulative absolute deviation (CAD), $\sum_{t=1}^{t_{max}} |\hat{A}_0(t) - A_0(t)|$ to assess the overall performance across a wide range of time. For the variance estimator of the four weighted estimators, we set the covariance vector $\Sigma_{\mu,S(t)}$ and off-diagonal elements of $\Sigma_{\mu}$ to 0, since such information is usually not available in practice.

The results are similar for different values of $\xi$ in the censoring distribution. To save space here we only present results for $\xi = -5$, that is, censoring is negatively correlated with $Z_1.$ Table 1 summarizes the results under A1, that is, $\Lambda_0(t)$ is consistent to $\Lambda_0(t)$ for the target cohort. Indeed, there is little bias for $\tilde{A}_0(t)$ whether or not the target cohort covariates are matched with those of the target cohort (C1–C4). On the other hand, the unweighted estimator $\hat{A}_0(t)$ that assumes $f(Z_1, Z_2) = f(Z_1, Z_2)$ is unbiased under C1, that is, $f^*(Z_1, Z_2) = f(Z_1, Z_2)$, and can be substantially biased when $f(Z_1, Z_2) \neq f(Z_1, Z_2)$ under C2–C4 with relative bias as high as 73%, even though $\Lambda_0(t)$ is $\Lambda_0(t)$. The weighted estimator, especially $\Lambda_0^{W_0}(t)$, has little bias with maximum relative bias 3% and the CAD and sMSE are often the lowest. The coverage probabilities of 95% confidence intervals are close to 95%. As expected, as more constraints are added, the bias of the weighted estimator generally becomes smaller. When $f^*(Z_1, Z_2) = f(Z_1, Z_2)$, both the unweighted and weighted estimators that leverage the (precise) summary information from target cohort improve the efficiency substantially with efficiency gain over 200%. Even when $f(Z_1, Z_2)$ is significantly different from $f^*(Z_1, Z_2)$ under C4, the weighted estimators do not lose much efficiency compared to the Breslow estimator. The asymptotic-based variance estimator is close
Table 2. Summary statistics of simulation results for Scenario A2: $\Lambda_0(t) = (0.01)t^2$, $\Lambda_1(t) = (0.01)t^{1.5}$, $n = 1000$, and $m = 100,000$.

| $t$ | $\Lambda_0(t)$ | $\Lambda_{0b}(t)$ | $\Lambda_{0w}(t)$ | $\Lambda_{0w}(t)$ | $\Lambda_{0w}(t)$ | $\Lambda_{0w}(t)$ |
|-----|----------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| 20  | 0.04           | PBias             | ESD               | ASE               | sMSE              | CP                |
|     | 1.10           | 0.37              | 0.37              | 0.37              | 0.37              | 0.35              |
|     | 1.10           | 0.43              | 0.30              | 0.30              | 0.30              | 0.36              |
|     | 1.10           | 0.30              | 0.30              | 0.37              | 0.37              | 0.36              |
|     | 0.0%           | 93.7%             | 93.7%             | 93.7%             | 93.7%             | 93.7%             |
| 40  | 0.16           | PBias             | ESD               | ASE               | sMSE              | CP                |
|     | 55.2%          | 0.1%              | 0.9%              | 0.9%              | 0.9%              | 0.9%              |
|     | 2.52           | 1.27              | 1.26              | 1.26              | 1.26              | 1.26              |
|     | 2.52           | 1.47              | 1.38              | 1.27              | 1.26              | 1.24              |
|     | 9.39           | 1.28              | 1.27              | 1.27              | 1.23              | 1.22              |
|     | 1.8%           | 97.9%             | 96.9%             | 97.5%             | 95.5%             | 95.0%             |
| 60  | 0.37           | PBias             | ESD               | ASE               | sMSE              | CP                |
|     | 25.5%          | 0.2%              | 0.8%              | 0.8%              | 0.8%              | 0.8%              |
|     | 4.83           | 2.49              | 2.47              | 2.34              | 2.34              | 2.34              |
|     | 4.83           | 2.83              | 2.61              | 2.41              | 2.41              | 2.39              |
|     | 10.41          | 2.51              | 2.48              | 2.36              | 2.36              | 2.36              |
|     | 53.7%          | 97.5%             | 96.3%             | 95.2%             | 95.4%             | 95.1%             |

CAD 3.89 0.46 0.45 0.44 0.43 0.43

| $t$ | $\Lambda_0(t)$ | $\Lambda_{0b}(t)$ | $\Lambda_{0w}(t)$ | $\Lambda_{0w}(t)$ | $\Lambda_{0w}(t)$ | $\Lambda_{0w}(t)$ |
|-----|----------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| 20  | 0.04           | PBias             | ESD               | ASE               | sMSE              | CP                |
|     | 1.07           | 0.44              | 0.44              | 0.42              | 0.42              | 0.46              |
|     | 1.10           | 0.51              | 0.48              | 0.45              | 0.44              | 0.47              |
|     | 9.35           | 1.02              | 1.02              | 1.01              | 0.76              | 0.46              |
|     | 0.0%           | 49.4%             | 53.8%             | 46.9%             | 73.0%             | 94.8%             |
| 40  | 0.16           | PBias             | ESD               | ASE               | sMSE              | CP                |
|     | 54.9%          | 11.8%             | 11.8%             | 11.8%             | 7.9%              | 7.9%              |
|     | 2.49           | 1.41              | 1.40              | 1.33              | 1.31              | 1.42              |
|     | 2.45           | 1.61              | 1.51              | 1.39              | 1.36              | 1.44              |
|     | 9.35           | 2.40              | 2.40              | 2.35              | 1.84              | 1.84              |
|     | 2.4%           | 82.6%             | 79.9%             | 75.9%             | 86.4%             | 95.7%             |
| 60  | 0.37           | PBias             | ESD               | ASE               | sMSE              | CP                |
|     | 25.0%          | 2.7%              | 2.7%              | 2.7%              | 1.2%              | 1.6%              |
|     | 4.87           | 2.56              | 2.53              | 2.37              | 2.35              | 2.43              |
|     | 4.87           | 2.87              | 2.65              | 2.45              | 2.43              | 2.50              |
|     | 10.37          | 2.74              | 2.72              | 2.56              | 2.39              | 2.51              |
|     | 53.4%          | 95.2%             | 95.2%             | 95.4%             | 95.9%             | 95.2%             |

CAD 3.87 0.82 0.82 0.80 0.62 0.49

Note: PBias, the relative bias; ESD, the empirical standard deviation ($\times 100$) of the 2000 estimates; ASE, the mean of the asymptotic-based standard error ($\times 100$); sMSE, the square root of mean squared error ($\times 100$); CP, the coverage probability of a 95% confidence interval for $\Lambda_0(t)$; CAD, the average cumulative absolute deviation between the estimated and true cumulative hazard function across a time range of one through sixty.

Table 2 shows the results when $\Lambda_0(t) \neq \Lambda_1(t)$ under A2 where they differ by the shape parameter. The results for the other two Scenarios A3 (different scale parameter) and A4 (different scale and shape parameters) have similar patterns and are provided in Supplemental Tables S2 and S3. Now $\Lambda_{0b}(t)$ has poor performance with relative bias ranging from 24.9% to 120.4%. When $f(Z_1, Z_2) = f^*(Z_1, Z_2)$, both $\Lambda_{0b}(t)$ and $\Lambda_{0w}(t)$ have little bias, and including additional moment constraints on risk factors $\hat{\Lambda}_{0b}(t)$’s improve efficiency slightly. When $f(Z_1, Z_2) \neq f^*(Z_1, Z_2)$, $\hat{\Lambda}_{0w}(t)$ is substantially biased and the weighted estimator, for example, $\hat{\Lambda}_{0w}(t)$, is nearly unbiased.

We also examined the performance of all the estimators under various situations. When the sample sizes for obtaining auxiliary information from the target cohort are moderate ($m = 1000$, Supplemental Tables S4–S7) and low ($m = 200$, Supplemental Tables S8–S11): when $\Lambda_0(t) = \Lambda_1(t)$, all estimators are unbiased and both $\hat{\Lambda}_{0b}(t)$ and $\hat{\Lambda}_{0w}(t)$ have comparable efficiency to $\hat{\Lambda}_{0b}(t)$ when $m = 1000$, and are less efficient when $m = 200$; however, when $\Lambda_0(t) \neq \Lambda_1(t)$ and $f(Z_1, Z_2) \neq f^*(Z_1, Z_2)$, even with $m = 200$, the auxiliary information can reduce the bias substantially with the weighted estimators (Table S9–S11). Other than moments, we considered quantile constraints. The moment constrained estimators have similar performance to the estimators with both moment and quantile constraints, and they both perform slightly better than the quantile constrained estimators (Supplemental Tables S12 and S13).

Finally, we considered the impact of competing risks on the proposed estimators (Web Appendix J). If the risk score
Table 3. Summary statistics of risk factors and outcome for WHI, PLCO, and UKB, and hazard ratio estimates (95% confidence interval [CI]) for CRC based on WHI.

| Risk factor/outcome | WHI | PLCO ctrl | PLCO inv | UKB England | UKB Scotland | UKB Wales | WHI HR (95% CI) |
|---------------------|-----|-----------|----------|-------------|-------------|-----------|----------------|
| No. of subjects     | 76,733 | 33,510 | 33,744 | 175,248 | 14,740 | 8070 | |
| Endoscopy history   | 56.2% | 10.1% | 73.9% | 37.7% | 24.7% | 23.7% | 0.79 (0.70, 0.89) |
| No. of relatives with CRC (≥1) | 14.7% | 11.1% | 11.5% | 11.9% | 12.4% | 10.7% | 1.24 (1.06, 1.45) |
| Vigorous leisure exercise 0-2 hr/wk | 14.9% | 23.3% | 23.1% | 33.9% | 33.6% | 30.9% | 0.99 (0.83, 1.18) |
| Vigorous leisure exercise >2 hr/wk | 12.1% | 11.6% | 11.6% | 14.5% | 13.6% | 13.0% | 0.83 (0.68, 1.03) |
| Aspirin/NSAID use (Regular user) | 85.0% | 47.6% | 48.0% | 24.6% | 26.0% | 24.9% | 0.76 (0.65, 0.90) |
| Vegetable intake (≥median portion/day) | 53.0% | 50.2% | 50.7% | 46.4% | 45.3% | 44.6% | 0.94 (0.83, 1.06) |
| BMI, kg/m² (≥30) | 22.8% | 23.6% | 24.1% | 24.1% | 24.2% | 28.4% | 1.39 (1.21, 1.59) |
| Estrogen status (Positive) | 44.9% | 50.6% | 50.7% | 8.3% | 6.8% | 7.6% | 0.87 (0.76, 0.99) |
| Mean follow-up years | 6.7 | 11.3 | 11.2 | 5.7 | 6.9 | 6.7 | |
| Other-cause mortality rate (5-year) | 3.8% | 2.0% | 2.1% | 1.4% | 1.6% | 1.4% | |
| CRC rate (5-year) | 0.7% | 0.5% | 0.3% | 0.5% | 0.5% | 0.4% | |
| No. of risk factors for constraints | 2 | 2 | 4 | 4 | 4 | |

NOTE: PLCO ctrl, PLCO control arm; PLCO inv, PLCO intervention arm. Gray context indicates the prevalence difference compared to WHI is greater than 10%.

Figure 3. Probabilities of developing CRC by age for the WHI and five target cohorts.

associated with the event of interest is also associated with the competing event, the proposed estimator is no longer unbiased; however, if such association is weak or the censoring proportion is moderate to high, the proposed estimators are nearly unbiased with the coverage probabilities of 95% CIs generally above 90%, reflecting good performance of the estimation and inference. Note that in a few settings imposing more moment-based constraints may increase the bias. This happens when the correlation of \( Z_1 \) and \( Z_2 \) differs greatly between the source and target populations, consistent with what we observe in the bivariate discrete variable situation in Section 3.1. However, the relative bias change is nearly negligible (<2%), implying that incorporating more constraints will not increase the bias in a noticeable way.

5. Application

In this section, we illustrate our methodology by recalibrating a CRC risk prediction model, established from the WHI, to five target cohorts: the intervention and control arms of the PLCO randomized trials and three countries (England, Scotland, and Wales) from the long-term UKB. For more details of these studies see Web Appendix K. In particular, we will use the WHI individual-level data to build the risk prediction model, and leverage the summary statistics from the target cohorts, which include overall hazard rates of CRC and means of key risk factors, to recalibrate the CRC baseline hazard function and calculate the t-year absolute risks for individuals in the target cohorts, accounting for the death from other causes as the competing event. We included eight risk factors based on the risk prediction model developed by Freedman et al. (2009) (Table 3). We fit the Cox proportional hazards model for women starting at age 50, as no women below age 50 were recruited in WHI.

Figure 3 shows the probability of developing colorectal cancer from 50 to 80 years old for the WHI and five target cohorts. It is obvious that these cohorts have very different disease probabilities, justifying the need to leverage CRC risk estimates from the target cohorts for recalibration. Accordingly, we incorporated the overall incidence rates at ages \( t = 50, 51, \ldots, 85 \) for both event of interest (CRC) and competing risk (non-CRC death) when recalibrating the WHI model.

Table 3 presents the descriptive statistics and hazard ratio estimates of risk factors. Generally, having a positive family history in the first-degree relatives and being obese increase CRC risk, while having had endoscopy, more exercise, Aspirin/NSAIDS use, greater vegetable intake and positive
estrogen status reduce risk. The prevalences of risk factors are similar between WHI and the two arms of PLCO, except for endoscopy history (WHI 56.2%, PLCO intervention arm 73.9%, PLCO control arm 10.1%) and Aspirin/NSAID use (WHI 85.0%, PLCO ~48%). For the three target cohorts from UKB, notably fewer participants have had endoscopy, use NSAIDS and have positive estrogen status but more participants have moderate exercise (0–2 hr/week) than WHI participants. Such discrepancy justifies the need to reweigh the risk factor distribution for the target PLCO and UKB cohorts.

A key aspect of model validation is that the model can predict the risk for developing disease accurately. To investigate the calibration performance, for each participant in the targeted cohort, we calculated the expected 5-year absolute risk of developing CRC from the enrollment age based on the risk prediction model, accounting for the competing risks of death from other causes. We stratified participants evenly into five groups according to their model-based risk estimates. Since participants entered the study at different ages, different estimates of \( \Lambda_0(t) \) can result in placing the participants into different risk strata. For each group, we calculated the empirical 5-year CRC risk accounting for competing risks using the nonparametric estimator proposed by Gray et al. (1988). A well calibrated model would yield risk estimates matching well to the empirical risks across all groups.

Figure 4 shows the comparison of average model-based and empirical absolute risk estimates in each risk group, for three models: the WHI model and the recalibrated models with unweighted \( \hat{\Lambda}^u_0(t) \) and weighted \( \hat{\Lambda}^w_0(t) \) (constraints for selected risk factors for each target cohort are highlighted in gray in Table 3). The WHI model used the competing risk hazard function estimated from the WHI data and the recalibrated models used the competing hazards based on the target cohort data. The WHI model is not well calibrated to both PLCO cohorts with model-based risk estimates far above the upper bounds of the 95% CI of the empirical estimates. Both recalibrated unweighted and weighted estimators improve the calibration substantially, with the 95% CIs of empirical estimates generally covering the model-based estimates. We calculated two summary measures of calibration: average absolute difference between model-based and empirical absolute risk estimates across five risk groups, and O/E ratio defined as the overall empirical 5-year absolute risk divided by the model-based 5-year absolute risk. The weighted estimator has smaller absolute differences (0.09% and 0.07%) and better O/E ratios (1.08 and 1.00) in both PLCO cohorts than the unweighted estimator (absolute difference, 0.13% and 0.09%; O/E, 0.86 and 0.94). As the PLCO cohort has a longer follow-up, we also calculated 10-year risk and the proposed weighted estimator has a very good calibration (Supplementary Figure S3). For the UKB cohorts, the weighted estimator again performs the best among the three models, with average absolute differences of 0.05%, 0.03%, and 0.09%, and O/E ratios of 1.02, 1.03, and 0.89 for England, Scotland, and Wales, respectively. The performance of the unweighted estimator is in between the WHI and weighted estimators. For the UKB England cohort, the unweighted estimator is poorly calibrated with the estimates not in the 95% CIs of empirical absolute risk in the three higher risk groups and the deviation is substantial in the highest risk group. This highlights the importance of accounting for the potential risk factor distribution difference when recalibrating the model. These results demonstrate that the weighted estimator leveraging the summary information about both the disease incidence and risk factors has the best performance.

To validate the performance of various methods in this data analysis, we investigated in-depth the performance of various methods and conducted a real-data based simulation study by simulating the outcomes using parameter estimates obtained from the real data, including competing risks modeling. The proposed estimator has comparable bias and efficiency to the Breslow estimator should the individual level data be available from the target data and is also robust to competing risk assumptions (results are shown in Web Appendix K).

6. Discussion

In this article, we present an empirical likelihood-based weighted estimating equation approach to recalibrate the risk prediction model developed from one cohort to another, leveraging the summary-level information from the target population. In the presence of competing risks, the proposed weighted estimating equation yields nearly unbiased estimator and valid inference, even if the risk score \( Z \) for the event of interest is associated with the competing event. The proposed method offers a practical and robust solution to recalibration of baseline risk while accommodating competing risks. While our example cohorts of PLCO and UKB have sufficient individual level data to build its own prediction model, in practice such cohorts from target populations often do not exist and recalibration with high-level information is a much more feasible and appealing solution than collecting individual-level data.

Huang, Qin, and Tsai (2016) proposed a novel double empirical likelihood-based approach to improving efficiency of parameter estimation by synthesizing information from the individual-level data with \( t \)-year survival probabilities from external sources. Here, we used the survival probabilities from the target population to recalibrate \( \Lambda_0(t) \) so that the projected risk estimation would not over- or under-estimate the true risks of patients in the target population. While efficiency gain would be an added bonus, our main objective is to reduce bias of recalibrated \( \Lambda_0(t) \). To achieve this goal, we used the empirical likelihood with the constraints of summary statistics of covariates from the target population to obtain weights such that the reweighted covariate distribution of the source study is closer to the target population. As a result, the key estimating equations (6) and theoretical development are fundamentally different from those in Huang, Qin, and Tsai (2016). Lastly, while Huang, Qin, and Tsai (2016) assumed both hazard ratios and overall hazards are transportable between the internal and external cohorts, our approach only requires hazard ratios to be the same.

An underlying assumption in our recalibration method is that the covariate effects are same between the source cohort and target population. While this assumption is generally considered acceptable in practice, one must carefully weigh this assumption based on domain knowledge and existing epidemiological evidence when comparing the effects across various cohorts and
populations. If there are individual level data from the target population, a formal testing may be conducted for the regression coefficients. However, such individual data are generally not available when we recalibrate the model for target populations, as otherwise there is no need to recalibrate the model. Some post-hoc analyses could potentially be done, when sufficient data have been collected in the target population after employing the recalibrated model.
Our proposed method can be extended in several directions. First, we imposed equality constraints in the empirical likelihood approach for reweighting the samples. When there is large uncertainty about the auxiliary information, our approach, though reducing the bias substantially, can lose efficiency. Inequality constraints may allow for some trade-off between bias and efficiency. Second, alternative approaches such as exponential tilting and the generalized empirical likelihood may generate weights with different properties (Little and Wu 1991). Third, the proposed method can be generalized to incorporate time-dependent predictors. Finally, additional subgroup survival probabilities can potentially help refine the recalibration and improve efficiency. These extensions warrant future investigations.

**Supplementary Materials**

The online supplementary materials contain the appendices for the article.

**Acknowledgments**

The authors gratefully acknowledge two referees, an associate editor, and the co-editors for their valuable comments and suggestions, which have significantly improved the article.

**Funding**

The work is supported in part by the grants from the National Institutes of Health (R01 CA189532, R01 CA195789, R01 CA236558 and U01CA86568) and Scientific Computing Infrastructure at Fred Hutch funded by ORIP grant SI00D028685. The authors are grateful to the generosity of WHI investigators for allowing them to use the WHI data to illustrate the method proposed in this article. The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN26820160001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C. The list of investigators is provided in Web Appendix F of the supplementary materials. Part of this research has been conducted using the UK Biobank Resource under Application Number 8614.

**References**

Abbasi, A., Peelen, L. M., Corpeleijn, E., van der Schouw, Y. T., Stolk, R. P., Spijkerman, A. M., van der A, D. L., Moons, K. G., Navis, G., Bakker, S. J., and Beulens, J. W. (2012), "Prediction Models for Risk of Developing Type 2 Diabetes: Systematic Literature Search and Independent External Validation Study," BMJ, 345, c5900. [2043]

Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P., McGinn, T., and Guyatt, G. (2017), "Discrimination and Calibration of Clinical Prediction Models: Users’ Guides to the Medical Literature," JAMA, 318, 1377–1384. [2043]

Breslow, N. (1972), "Regression Models and Life-Tables," Journal of the Royal Statistical Society, Series B, 34, 187–222. [2043]

Cox, D. R. (1972), "Regression Models and Life-Tables," Journal of the Royal Statistical Society, Series B, 34, 187–222. [2043]

Drost, F.-J. H., Nieboer, D., Morgan, T. M., Carroll, P. R., Roobol, M. J., Trock, B., Ehdaiie, B., Carroll, P., Filson, C., Kim, J., and Logothetis, C. (2019), "Predicting Biopsy Outcomes During Active Surveillance for Prostate Cancer: External Validation of the Canary Prostate Active Surveillance Study Risk Calculators in Five Large Active Surveillance Cohorts," European Urology, 76, 693–702. [2043]

Freedman, A. N., Slattery, M. L., Ballard-Barbash, R., Willis, G., Cann, B. J., Pee, D., Gail, M. H., and Pfeiffer, R. M. (2009), "Colon/Rectal Cancer Risk Prediction Tool for White Men and Women Without Known Susceptibility," Journal of Clinical Oncology, 27, 686–693. [2052]

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. (1989), "Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually," Journal of the National Cancer Institute, 81, 1879–1886. [2043,2044]

Gray, R. J. (1988), "A Class of k-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk," The Annals of Statistics, 16, 1141–1154. [2053]

Huang, C.-Y., Qin, J., and Tsai, H.-T. (2016), "Efficient Estimation of the Cox Model With Auxiliary Subgroup Survival Information," Journal of the American Statistical Association, 111, 787–799. [2044,2053]

Keiding, N., and Louis, T. A. (2016), "Perils and Potentials of Self-Selected Entry to Epidemiological Studies and Surveys," Journal of the Royal Statistical Society, Series A, 179, 319–376. [2043,2045]

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," The Annals of Mathematical Statistics, 22, 79–86. [2047]

Little, R. J., and Wu, M.-M. (1991), "Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ," Journal of the American Statistical Association, 86, 87–95. [2055]

Liu, D., Zheng, Y., Prentice, R. L., and Hsu, L. (2014), "Estimating Risk With Time-to-Event Data: An Application to the Women’s Health Initiative," Journal of the American Statistical Association, 109, 514–524. [2043,2045]

Owen, A. B. (2001), Empirical likelihood, Boca Raton, FL: Chapman and Hall/CRC. [2044,2045]

Pfeiffer, R. M., and Gail, M. H. (2017), Absolute Risk: Methods and Applications in Clinical Management and Public Health, Boca Raton, FL: Chapman and Hall/CRC. [2045,2046]

Prorok, P. C., Andriele, G. L., Bresalier, R. S., Buys, S. S., Chia, D., Crawford, E. D., Fogel, R., Gelman, E. P., Gilbert, E. H., Hasson, M. A., and Hayes, R. B. (2000), "Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial," Controlled Clinical Trials, 21, 2735–3095. [2044]

Qin, J., and Lawless, J. (1994), "Empirical Likelihood and General Estimating Equations," The Annals of Statistics, 22, 300–325. [2044,2045]

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010), "Assessing the Performance of Prediction Models: A Framework for Some Traditional and Novel Measures," Epidemiology, 21, 128–138. [2043]

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., and Liu, B. (2015), "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age," PLoS Med, 12, e1001779. [2044]

Van Calster, B., and Vickers, A. J. (2015), "Calibration of Risk Prediction Models: Impact on Decision-Analytic Performance," Medical Decision Making, 35, 162–169. [2043]

The Women’s Health Initiative Study Group (1998), "Design of the Women’s Health Initiative Clinical Trial and Observational Study," Controlled Clinical Trials, 19, 61–109. [2044]

Zhao, W., Zheng, J., Chen, Y. Q., and Hsu, L. (2020), "Adjusted Time-Varying Population Attributable Hazard in Case–Control Studies," Statistical Methods in Medical Research, 29, 243–257. [2045]