# Estimating restricted mean survival time and expected life-years lost in the presence of competing risks within flexible parametric survival models

**Sarwar Islam Mozumder** ( sarwar.islam@gmail.com )
University of Leicester https://orcid.org/0000-0001-9644-7525
**Mark Rutherford**
University of Leicester College of Medicine Biological Sciences and Psychology
**Paul Lambert**
University of Leicester College of Medicine Biological Sciences and Psychology

**Research article**

**Keywords:** competing risks, restricted mean survival time, restricted mean life time, flexible parametric model, life-years lost, survival analysis

**Posted Date:** January 7th, 2021

**DOI:** https://doi.org/10.21203/rs.2.23839/v3



**Version of Record:** A version of this preprint was published on March 11th, 2021. See the published version at https://doi.org/10.1186/s12874-021-01213-0.

RESEARCH

# Estimating restricted mean survival time and expected life-years lost in the presence of competing risks within flexible parametric survival models

Sarwar I Mozumder[1*], Mark J Rutherford[1] and Paul C Lambert[1,2]

[*]Correspondence:
sarwar.mozumder@le.ac.uk
[1]Biostatistics Research Group, Department of Health Sciences, University of Leicester, University Road, LE1 7RH Leicester, UK
Full list of author information is available at the end of the article

**Abstract**

**Background:** Royston-Parmar flexible parametric survival models (FPMs) can be fitted on either the cause-specific hazards or cumulative incidence scale in the presence of competing risks. An advantage of modelling within this framework for competing risks data is the ease at which alternative predictions to the (cause-specific or subdistribution) hazard ratio can be obtained. Restricted mean survival time (RMST), or restricted mean failure time (RMFT) on the mortality scale, is one such measure. This has an attractive interpretation, especially when the proportionality assumption is violated. Compared to similar measures, fewer assumptions are required and it does not require extrapolation. Furthermore, one can easily obtain the expected number of life-years lost, or gained, due to a particular cause of death, which is a further useful prognostic measure as introduced by Andersen.

**Methods:** In the presence of competing risks, prediction of RMFT and the expected life-years lost due to a cause of death are presented using Royston-Parmar FPMs. These can be predicted for a specific covariate pattern to facilitate interpretation in observational studies at the individual level, or at the population-level using standardisation to obtain marginal measures. Predictions are illustrated using English colorectal data and are obtained using the Stata post-estimation command, standsurv.

**Results:** Reporting such measures facilitate interpretation of a competing risks analysis, particularly when the proportional hazards assumption is not appropriate. Standardisation provides a useful way to obtain marginal estimates to make absolute comparisons between two covariate groups. Predictions can be made at various time-points and presented visually for each cause of death to better understand the overall impact of different covariate groups.

**Conclusions:** We describe estimation of RMFT, and expected life-years lost partitioned by each competing cause of death after fitting a single FPM on either the log-cumulative subdistribution, or cause-specific hazards scale. These can be used to facilitate interpretation of a competing risks analysis when the proportionality assumption is in doubt.

**Keywords:** competing risks; restricted mean survival time; restricted mean life time; flexible parametric model; life-years lost; survival analysis

## 1 Background

In observational studies of time-to-event data, researchers are often interested in decomposing the overall probability of death into component parts due to the event of interest, and competing, but mutually exclusive outcome events. For example, in cancer studies, it is of interest to partition the overall probability of death into the probability of death due to cancer and the probability of death due to other causes. These are referred to as cause-specific cumulative incidence functions (CIFs) and are often chosen as the primary estimand of interest. The cause-specific CIF gives the probability of dying from the cause of interest at a particular time whilst also being at risk of dying from other causes of death [1, 2]. In order to arrive at these quantities and to circumvent bias, methods that appropriately account for the competing nature of the events must be applied. The restricted mean failure time (RMFT) has been proposed as an alternative summary measure that is based on the area under the all-cause probability of death up to a specific time-point[3]. In an analogous way to the decomposition into cause-specific CIFs, the RMFT can be further partitioned to give the expected number of life years lost due to a specific cause before a given time-point. In this paper, we describe how the aforementioned measures can be obtained using a flexible parametric model (FPM) as the estimation approach by modelling covariate effects either using (1) the direct relationship with the cause-specific CIF on the subdistribution hazards (SDHs) scale, or (2) modelling all cause-specific hazard functions (CSHs) to obtain each cause-specific CIF [4, 5, 6, 7]. Choosing FPMs as the estimation method allows us to estimate effects conditional on covariates, and effects averaged over specific covariate distributions.

Forming contrasts to compare exposure groups is often a further key focus in many large population-based studies. A common approach would be to report either cause-specific hazard ratios (HRs), which measures the effect of an exposure group on the rate of dying from a cause of interest, or sub-distribution hazard ratios (SHRs), which measures the effect of an exposure group on the risk of dying from a cause

of interest, whilst assuming that the cause-specific HR or SHR was constant over time. However, it is well known, for instance, that the HR for tumor size in cancer studies will vary over time since diagnosis, with stronger relative effects shortly after diagnosis [8, 9, 10]. When non-proportional hazards are present i.e. when the HR is expected to change over time, it has been argued that the HR as the target estimand is not appropriate and there are further issues in making causal inferences using HR measures due to its non-collapsibility as a relative risk measure [11]. As an alternative to the HR, estimation of the difference in restricted mean survival time (RMST), also known as the restricted mean lifetime (RMLT), as the primary estimand has been proposed [12, 13, 14, 15, 16, 17, 18, 19]. This, in contrast to the HR, is known as a collapsible measure [11, 20]. Furthermore, this single summary measure can still be presented when relaxing the assumption of proportional hazards within the model-building process. These can either be presented as conditional differences, which is the average covariate effect on the individual, or marginal differences, which is the average covariate effect on the population [21].

In the presence of competing risks, Andersen [3] introduces the analogue to the RMST measure for the CIF which gives the (total) number of years lost before a pre-specified time, i.e. RMFT, and demonstrates how this can be partitioned to give the expected number of life-years lost due to each cause of death [22]. In his approach, he estimates RMFT and expected number of life-years lost using regression models with pseudo-observations [3, 23]. These models only allow prediction for specific quantities of interest and only at single time-points. Therefore separate models must be fitted to estimate, for example, either the cause-specific CIF or RMFT, when it may be of interest to obtain both and at various time-points. For instance, to allow comparability and to obtain the entire picture of the impact of different groups on outcome, it has been suggested that differences in RMST, RMFT and therefore, expected number of life-years lost, should be reported along-

side their respective survival, or cumulative incidence functions [24]. Alternatively, the Royston-Parmar FPM approach for estimating RMST, which is extended for competing risks to estimate partitioned RMFT, as introduced in this paper, can be used [25]. In contrast to more popular approaches, such as the Cox model, a parametric estimate of the baseline hazard function is obtained as part of the full likelihood function. This is estimated using restricted cubic splines (RCS), allowing easy prediction of absolute comparisons between key quantities of interest. What's more, standard errors for predictions can be estimated via the delta method, which offers computational advantages in larger data compared to approaches for non-parametric and semi-parametric methods which use bootstrapping, or jack-knife resampling methods [26]. Further advantages include the easy inclusion of time-dependent effects using interactions with RCS for relaxing the proportional hazards assumption. Estimating both the baseline effects, and time-dependent effects to model departures from the baseline using splines allows a unified approach for estimating all required parameters in order to obtain predictions of all quantities of interest. Therefore, we introduce in this paper how RMFT as the chosen estimand can be estimated using FPMs in the presence of competing risks on either the CSHs or cumulative incidence scale as the estimator [7, 5]. This extends on previous work by Royston and Parmar where estimation in the presence of competing risks is not considered [16]. This approach allows the researcher to obtain differences in effect between exposure groups either conditional on a set of covariates, or averaged over a covariate distribution, also known as marginal estimates. Furthermore, both marginal and conditional estimates can be obtained from the same model where the prediction of marginal estimates using standardisation is proposed [27, 28]. We, therefore, further demonstrate how difference in marginal estimates of RMFT as the chosen estimand for the comparison between covariate groups can be obtained within FPMs for competing risks.

We begin with a brief review of competing risks in Section 2.1 and highlight particular interest in the cause-specific CIF. This is followed by an introduction of the RMFT as the chosen estimand in Section 2.2 along with other useful measure such as expected life-years lost. Section 2.3 details FPM approaches for estimation in the presence of competing risks. In Section 2.4, we show how absolute differences between RMFT and expected number of life-years lost are calculated to assess the impact of a covariate. We further demonstrate how these models can be used for easily obtaining marginal estimates and associated contrasts using standardisation in Section 2.4.4. For illustration of these various measures, English colorectal cancer data obtained from National Cancer Registration and Analysis Service (NCRAS) is analysed in Section 3 where comparisons between the most and least deprived colorectal cancer patients are made, accompanied by Stata code for estimation in Appendix D. Finally, the paper is concluded with a discussion on the use and estimation of RMST in the presence of competing risks within FPMs. Although we specifically consider application to cancer studies, where the event of interest is death from cancer, the methods are generalizable to other time-to-event data and therapeutic areas.

# 2 Methods

## 2.1 Overview of competing risks

In the presence of competing risks, an individual is at risk of failing from more than one event where the occurrence of one event means that others cannot occur. In the context of a cancer survival study, this is when a patient can die from a multitude of other causes as well as the cancer itself. However, if the patient dies from one of these other causes, it means that the time at which the patient would have died from cancer is never observed. One of the key quantities, and often the chosen estimand of interest within this framework, is the cause-specific CIF [1].

### *2.1.1 Cause-specific CIF*

Let $T$ be a non-negative random variable for the time to death from any cause. Furthermore, let $D$ denote the cause of death in the presence of $k = 1, \ldots, K$ competing risks, where $D = 1, \ldots, K$. It follows that the estimand, cause-specific CIF, $F_k(t)$, is defined as,

$$F_k(t) = P(T \leq t, D = k) \tag{1}$$

This is interpreted as the probability of dying from cause $k$ by time $t$ whilst also being at risk of dying from other competing causes of death. Note here that the cause-specific CIF is an improper distribution function since the integral of $F_k(t)$ at infinity is always less than 1 [3].

The target estimand, the cause-specific CIF, can be calculated using either all $k$ CSH functions, or by utilising the one-to-one relationship between the cause-specific SDH function. These are briefly introduced below.

### *2.1.2 Cause-specific hazards*

The CSHs, $h_k^{cs}(t)$, give the instantaneous mortality rate from a particular cause $k$ given that the patient is still alive at time $t$ in the presence of all the other causes of death such that,

$$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t, D = k | T \geq t)}{\Delta t} \tag{2}$$

It follows that the target estimand, the cause-specific CIF, can be calculated as a function of all $k$ CSH functions,

$$F_k(t) = \int_0^t S(u) h_k^{cs}(u) \mathrm{d}u \tag{3}$$

where $S(t) = \exp\left(-\sum_{k=1}^{K} \int_0^t h_k^{cs}(u)\mathrm{d}u\right)$ is the all-cause survival function.

### *2.1.3 Subdistribution hazards*

Alternatively, Gray [29] introduces the SDH function for cause $k$, $h_k^{sd}(t)$, which offers a direct one-to-one relationship with the cause-specific CIF estimand. This has the following mathematical formulation,

$$h_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t, D = k | T \geq t \cup (T \leq t \cap D \neq k)}{\Delta t} \tag{4}$$

which is interpreted as the instantaneous "sub"-rate of failure at time $t$ from cause $k$ amongst those who are still alive, or have died from any of the other $K-1$ competing causes excluding cause $k$ [30].

This is not defined as a typical epidemiological rate since the risk-set includes those that are either still alive *or* have died from a competing cause of death. However, if individuals do not experience the competing event, then the SDH rate and the CSH rate are both equivalent [31]. It should be noted that, due to the nature of the risk-set in the definition of a SDH, it is very difficult to interpret [32, 30, 33].

The cause-specific CIF estimand can be directly obtained from the SDH for cause $k$ using the standard survival transformation of the cumulative SDH function for cause $k$, $H_k^{sd}(t)$, such that,

$$F_k(t) = 1 - \exp\left[-H_k^{sd}(t)\right] \tag{5}$$

This shows that a one-to-one correspondence is maintained between the SDH function for a specific cause of death and the cause-specific CIF.

The choice of which scale to model on depends entirely on the research question to be answered which would relate to other quantities specific to the modelling approach that may be of interest. For instance, if primary interest is in aetiological outcome, then the estimand of interest would be on the CSH rates. For interest in prognostic outcome, one may wish to quantify effects on the risk of dying from a specific cause of death. In this case, the estimand of interest would be the cause-specific CIF, which can be obtained as function of all CSHs, or through the SDH for cause $k$. Further discussion on this topic is provided elsewhere [34, 4].

## 2.2 Overview of restricted mean survival time for competing risks

The RMST measure quantifies the average survival, or time lived, of a patient from time 0 up to a pre-defined time-point, $t^*$. In the absence of competing risks, the RMST before $t = t^*$, $\mu(t^*)$, of a random variable $T$ is equal to the expectation of $\min(T, t^*)$. RMST, in the absence of covariates, can be expressed as the estimand,

$$\mu(t^*) = E(\min(T, t^*)) = \int_0^{t^*} S(u)\mathrm{d}u \tag{6}$$

where $S(t)$ is the all-cause survival function. If time is measured in years, this is the average life-years lived before time $t^*$. The choice of $t^*$ should be pre-determined and clinically motivated, and will vary by, for example, cancer types [16, 15]. This is also often chosen at maximum follow-up time [13, 35].

In addition to this, Andersen [3] proposes calculation of the expected number of years lost before time $t^*$ such that the estimand can be defined as,

$$L(0,t^*) = t^* - E(\min(T,t^*)) = t^* - \int_0^{t^*} S(u)\mathrm{d}u \tag{7}$$

### *2.2.1 Expected loss in life due to a cause of death*

In the presence of competing risks, Andersen [3] shows that the (total) number of years lost, $L(0,t^*)$, can be decomposed into the number of years lost due to each cause $k$ [22]. It follows that since,

$$S(t) = 1 - \sum_{k=1}^{K} F_k(t) \tag{8}$$

then the RMST in Equation 6 can be expressed as a function of each cause-specific CIF through the following integral,

$$\begin{aligned}\mu(t^*) = E(\min(T,t^*)) &= \int_0^{t^*} S(u)\mathrm{d}u = \int_0^{t^*} 1 - \sum_{k=1}^{K} F_k(u)\mathrm{d}u \\ &= t^* - \int_0^{t^*} \sum_{k=1}^{K} F_k(u)\mathrm{d}u\end{aligned} \tag{9}$$

Equation 7 can also be written as a sum of the integral of each cause-specific CIF such that,

$$L(0,t^*) = t^* - \int_0^{t^*} S(u)\mathrm{d}u = \sum_{k=1}^{K} \int_0^{t^*} F_k(u)\mathrm{d}u \tag{10}$$

which may also be referred to as restricted mean failure time (RMFT). It follows that RMFT can be partitioned where we have the estimand,

$$L_k(0, t^*) = \int_0^{t^*} F_k(u)\mathrm{d}u \tag{11}$$

which gives the expected number of years lost due to cause $k$ before time $t^*$.

## 2.3 Flexible parametric survival models

For competing risks data, many adopt the cause-specific Cox proportional hazards model, or the Fine & Gray approach as the chosen estimator for the estimands introduced in Sections 2.1 and 2.2. Here, we propose the use of FPMs as the chosen estimator in order to obtain the estimand of interest. FPMs are increasing in popularity since the baseline SDH or CSH function is estimated as part of a fully specified likelihood function and allows the estimation of various estimands from a single model [5, 7]. These models were introduced for standard survival data (in the absence of competing risks) on various scales by Royston and Parmar [9] using a general link function, $g(\cdot)$, to better capture and represent the behaviour of real world data. To increase flexibility and more accurately capture complex shapes of the cumulative hazard function, Royston and Parmar [9] proposed the use of RCS (see Appendix A). Under the assumption of proportional hazards, Rutherford et. al [36] showed in simulations that FPMs more accurately capture complex shapes of hazard functions. They further illustrated that unbiased estimates of the HRs were obtained. Given a vector of $M$ knots, $\mathbf{m}$, and a vector of $M-1$ parameters, $\boldsymbol{\gamma}$, with a RCS function, $s(\ln(t); \boldsymbol{\gamma}, \mathbf{m})$ we have that,

$$\eta = g(G_k(t \mid \mathbf{x}_k)) = s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k \boldsymbol{\beta}_k^T \tag{12}$$

where, $\beta$, is a vector of co-efficient parameters and, $\mathbf{x}$, is a vector of covariates.

Equation 12 can also be easily extended for time-dependent effects to model non-proportionality by fitting interactions between the associated covariates and the spline functions. Using this interaction, a new set of knots, $\mathbf{m}_e$, are introduced, which represent the $e^{th}$ time-dependent effect with associated parameters $\boldsymbol{\alpha}_e$. If there are $e = 1, \cdots, E$ time-dependent effects, Equation 12 can be extended such that,

$$\eta = g(G_k(t \mid \mathbf{x})) = s(\ln(t); \boldsymbol{\gamma}, \mathbf{m}_0) + \mathbf{x}\boldsymbol{\beta}^T + \sum_{l=1}^{E} s(\ln(t); \boldsymbol{\alpha}_l, \mathbf{m}_l) x_l \tag{13}$$

Non-proportional hazards are a common occurrence in studies with long follow-up time, or, in the context of cancer studies, when the effect of covariates (e.g tumor size, or treatment) on cancer-related mortality varies over time [9, 10, 8, 19]. FPMs, extended for time-dependent effects as in the Equation above, have also been shown to accurately capture complex shapes of the hazard function with time-dependent effects i.e. where there is non-proportionality in the hazards [37]. This result is consistent with what was shown by Rutherford et. al. for FPMs without time-dependent effects i.e. proportional hazards, as mentioned above [36]. Further technical details on FPMs for standard survival data in the absence of competing risks can be found elsewhere [9, 38, 25].

The models described in Equations 12 and 13 can be fitted on either CSHs scale [7], where $G_k(t \mid \mathbf{x}) = S_k(t \mid \mathbf{x})$, or cumulative incidence scale [5, 6], where $G_k(t \mid \mathbf{x}) = 1 - F_k(t \mid \mathbf{x})$, based on different link functions, $g(\cdot)$. The relationship of these with the cause-specific CIF are defined in Sections 2.1.2 and 2.1.3. Therefore, it follows that, using a complementary log-log link function, the corresponding log-cumulative CSHs FPM (otherwise referred to as a cause-specific FPM), is,

$$\eta_k^{cs} = \log\left[-\log\left[S_k\left(t \mid \mathbf{x}_k\right)\right]\right] = \log\left[H_k^{cs}\left(t \mid \mathbf{x}\right)\right] = s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\boldsymbol{\beta}_k^T \quad (14)$$

and can be fitted in a similar way to the standard FPM. Alternatively, models for all $k$ causes can be fitted simultaneously by restructuring the data as described by Hinchliffe et. al. [7].

The log-cumulative SDHs FPM for cause $k$ (also known as the flexible parametric cumulative incidence model, or FPCIM), on the other hand is defined as,

$$\eta_k^{sd} = \log\left[-\log\left[1 - F_k\left(t \mid \mathbf{x}_k\right)\right]\right] = \log\left[H_k^{sd}\left(t \mid \mathbf{x}\right)\right] = s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\boldsymbol{\beta}_k^T \quad (15)$$

and can be fitted using the approach outlined using either the full likelihood function as described by Mozumder et. al. [5] or by using time-dependent censoring weights, similar to the Fine-Gray model, as detailed by Lambert et. al. [6]. As previously mentioned, alternative link functions are also available for models on either scale. See for example, Lambert et. al. [6].

## 2.4 Estimation

### *2.4.1 Cause-specific cumulative incidence function*

If modelling on the cumulative incidence scale using SDHs, after fitting the FPCIM in Equation 15, the cause-specific CIF is obtained by the following,

$$\widehat{F}_k\left(t \mid \mathbf{x}\right) = 1 - \exp\left(-\exp\left(\hat{\eta_k^{sd}}\left(t \mid \mathbf{x}\right)\right)\right) \quad (16)$$

Alternatively, when modelling on the CSHs scale, after fitting the cause-specific FPM in Equation 14, and as shown in Equation 3, the integral below must be evaluated in order to obtain the cause-specific CIF,

$$\widehat{F}_k\left(t \mid \mathbf{x}\right) = \int_0^t \widehat{S}\left(u \mid \mathbf{x}\right) \widehat{h}_k^{cs}\left(u \mid \mathbf{x}\right) du \tag{17}$$

where the predicted CSH function is,

$$\widehat{h}_k^{cs}\left(t \mid \mathbf{x}\right) = \frac{ds\left(\log(t) \mid \gamma, \mathbf{m}_0\right)}{dt} \exp\left(\eta_k^{cs}(t)\right) \tag{18}$$

and the predicted all-cause survival function is,

$$\widehat{S}\left(u \mid \mathbf{x}\right) = \prod_{k=1}^{K} \exp\left(-\int_0^t \widehat{h}_k^{cs}(u \mid \mathbf{x}) \mathrm{d}u\right) \tag{19}$$

However, as the above integral is not of closed form, numerical approximation techniques must be used. Here, the Gauss-Legendre quadrature approximation method is used [39]. Details of this method is provided in Appendix B. Therefore, after fitting the cause-specific FPM for each $k$ causes, the predicted cause-specific CIF at $t_1, \cdots, t$ different time-points over an interval $[0, t]$ is approximated by applying Gaussian quadrature rules with $W(u) = 1$ such that,

$$\widehat{F}_k\left(t \mid \mathbf{x}\right) = \int_0^t f_k^*(u) \mathrm{d}u \approx \frac{t-0}{2} \sum_{i=1}^{m} w_i' f_k^*\left(\frac{t-0}{2} u_i' + \frac{t+0}{2} \mid \mathbf{x}\right) \tag{20}$$

where, $\widehat{f}_k^*(t)$, is the "sub"-density function such that,

$$\widehat{f}_k^{*}\left(t \mid \mathbf{x}\right) = \widehat{S}\left(t \mid \mathbf{x}\right)\widehat{h}_k^{cs}\left(t \mid \mathbf{x}\right) \tag{21}$$

### *2.4.2 Restricted mean failure time and expected number of life-years lost due to each cause of death*

If RMFT is the chosen target estimand of interest, this can be predicted as the integral under the all-cause CIF such that,

$$\widehat{L}\left(0, t^{*}\right) = \int_0^{t^{*}} \sum_{j=1}^{K} \widehat{F}_j(u)\mathrm{d}u = \sum_{j=1}^{K} \int_0^{t^{*}} \widehat{F}_j(u)\mathrm{d}u \tag{22}$$

where the predicted expected number of life-years lost before time $t^{*}$ due to each cause $k$ is,

$$\widehat{L}_k\left(0, t^{*} \mid \mathbf{x}\right) = \int_0^{t^{*}} \widehat{F}_k(u \mid \mathbf{x})\mathrm{d}u \tag{23}$$

Again, as above in Equation 17, as the integral is of closed-form, we use the Gauss-Legendre quadrature approximation technique to numerically evaluate,

$$\int_0^{t^{*}} \widehat{F}_k(u)\mathrm{d}u \approx \frac{t^{*}-0}{2} \sum_{i=1}^{m} w_i' \widehat{F}_k\left(\frac{t^{*}-0}{2}u_i' + \frac{t^{*}+0}{2} \mid \mathbf{x}\right) \tag{24}$$

It follows that the RMST can also be obtained by,

$$\widehat{\mu}\left(t^* \mid \mathbf{x}\right) = t^* - \sum_{j=1}^{K} \widehat{L}_j\left(0, t^*\right) \tag{25}$$

### *2.4.3 Conditional differences*

In population-based studies, i.e. non-randomised studies, it may be of interest to make absolute or relative comparisons between different covariate groups. As an alternative summary measure, or estimand, to the HR, we can calculate the difference in RMST between two covariate groups, or the difference in expected loss in life due to different causes [19]. Let $X$ be a binary covariate that denote the group of interest and $Z$ be the set of measured covariates with a specific covariate pattern $\mathbf{z}_j$. To estimate the average number of life years gained in group $X = 0$ compared to group $X = 1$, we have that,

$$\hat{\mu}(t^* \mid X = 1, Z = \mathbf{z}_j) - \hat{\mu}(t^* \mid X = 0, Z = \mathbf{z}_j) \tag{26}$$

Alternatively, we can also estimate the expected reduction in the loss (or gain) in life due to cause $k$ by,

$$\hat{L}_k(0, t^* \mid X = 1, Z = \mathbf{z}_j) - \hat{L}_k(0, t^* \mid X = 0, Z = \mathbf{z}_j) \tag{27}$$

Partitioning in this way is particularly useful if covariates act differently on different causes of death. For example, those from a particular covariate group may lose (or gain) some life-years due to a specific cause of death in comparison to another covariate group.

Absolute measures of gains or losses in years of life are presented above as potential estimates of interest. To obtain relative measures, the ratio between the RMST estimates, or expected loss in life due to cause $k$ for the two covariate groups are calculated. Extension can also be made for comparisons on a unit increase in a continuous covariate $Z$, and for time-dependent effects.

### *2.4.4 Standardisation for marginal differences*

Regression standardisation is part of the estimator that can be used to obtain marginal predictions for different covariate groups at each observation given a set of measured confounders [28, 27]. Here, we apply standardisation to RMST and cause-specific CIFs estimates obtained from a flexible parametric competing risks survival model. In this case, it is of interest to compare the average life-years lived before time $t^*$ between two different groups [18, 17]. This is done by obtaining marginal estimates which are calculated as an average over every individual in the observed dataset. This enables comparisons that solely focus on the differences between the two groups of interest by forcing the same covariate distribution over multiple confounders. If all exposures and confounders are measured at baseline, this is essentially equivalent to the G-formula [40]. For example, to compare males and females, estimates must be standardised by age in order to force the same age distribution for both males and females. Extension can be made for multiple covariates and other potential confounders. This is calculated using an average of RMST estimates for each patient to summarise the risk for a certain covariate group. For instance, let $X$ be an indicator variable that denotes the group of interest and $Z$ be the set of measured covariates. Then the predicted RMST estimate for the $i^{th}$ individual, where $i = 1, \ldots, N$, is,

$$\widehat{\mu}_i = t^* - \int_0^{t^*} \sum_{k=1}^{K} \left[ \widehat{F}_k(u \mid X = x, Z = z_i) \right] \mathrm{d}u \tag{28}$$

where $X$ is fixed to a specific value, $x$, and $Z$ is the observed covariate pattern, $z_i$, for the $i^{th}$ individual. We can then average over the marginal distribution of $Z$ for all the predicted restricted mean life estimates obtained for each individual $i$ such that,

$$E(\widehat{\mu}^{stand} \mid X = x, Z) = \frac{1}{N} \sum_{i=1}^{N} \widehat{\mu}_i \tag{29}$$

This allows us to calculate marginal differences between covariate groups. For example, between group $X = 0$ and group $X = 1$, the marginal difference in RMST is,

$$E(\widehat{\mu}^{stand} \mid X = 1, Z) - E(\widehat{\mu}^{stand} \mid X = 0, Z) \tag{30}$$

In recent literature, some have advocated the use of RMST as a causal measure [41, 42]. For a causal interpretation, the consideration of additional assumptions are required and by adjusting for all appropriate confounders, these measures can be extended and interpreted as causal effects and thus, used as an estimand [21]. This is because, as shown above, they provide marginal comparisons averaged over the same covariate distribution by using standardisation. Standardisation, otherwise referred to as G-computation, has also been highlighted by Gran et al. [43] as an approach for obtaining useful summary causal-effect measures in more complicated multi-state models. However, this is beyond the scope of the paper and estimation of causal effects are not explicitly discussed here. Note also that we only consider time-fixed confounders and that there are additional complexities when considering time-dependent risk-groups [44].

# 3 Results: Colorectal cancer survival in England

## 3.1 Data

Data was obtained from the National Cancer Registration and Analysis Service (NCRAS) to illustrate the estimation of various measures introduced in Section 2.2. The data consist of English colorectal (ICD10: C18, C19 and C20) male and female cancer patients aged between 45 and 90 years old. Patients are diagnosed on or after 1998 are included with follow-up restricted to either 10 years or censored at 31 Dec 2013, whichever comes first. Analysis is further restricted to patients from the most or least deprived groups as defined by the upper and lower quintiles of the English index of multiple deprivation 2010 (IMD 2010). These groups are selected to simplify analysis and to make for easy illustration of presenting different metrics to allow comparisons between the two groups. The final data consisted a total of 159,022 individuals of which 48,845 die from cancer, 7,987 from cardiovascular disease (CVD) and 32,133 from other causes. In Appendix C, summary statistics on the age distribution, and number of patients in each deprivation and sex groups are provided.

## 3.2 Model

For demonstration purposes, predictions are obtained after fitting an FPCIM simultaneously for all $k$ causes of death and standard errors for confidence intervals (CIs) are obtained using the delta method. However, predictions are also available after fitting cause-specific FPMs. This paper focusses on the various estimands we can obtain from such models, namely, the RMST measure and expected life-years lost.

Models are fitted simultaneously for all $k$ causes of death using the approach of Lambert et al. [6] and Geskus [45]. This fits the model after restructuring the data and applying time-dependent weights that are obtained parametrically to the censoring distribution of the competing causes of death. Alternatively, using the

approach described by Jeong and Fine [46], models can be fitted on individual-level data using the full likelihood function [47]. Models for each of the causes of death include sex, IMD 2010 deprivation group (upper and lower quintile only) and a non-linear effect of continuous age using RCS with 3 DF centred at 45 years old at diagnosis. Time-dependent effects to relax the proportionality assumptions are included for sex, non-linear age and deprivation group with 2 DF and 3 DF are used for the baseline RCS function. In order to evaluate whether assuming non-proportional (subdistribution) hazards was more sensible, and is more consistent with the data, a likelihood ratio test was performed. This compared the FPCIM with time-dependent effects to relax the proportionality assumption to the one without that assumed proportional SDHs. The likelihood ratio test statistic was 752.94 and the associated p-value was less than 0.0001. This shows that relaxing the proportionality assumption leads to a statistically significant improvement in model fit. Note that this is an illustrative model and we therefore omit formal evalutation of the model performance. When evaluating the model in practice, we recommend conducting a sensitivity analysis, particularly in the selection of the number of knots. This can be done by comparing the Akaike information criterion and the Bayesian information criterion as an informal guide to selecting the appropriate number of knots and covariates [6].

## 3.3 Analysis of data with conditional estimates

### *3.3.1 Cause-specific cumulative incidence functions*

Cause-specific CIFs are presented in Figure 2 for male colorectal cancer patients. The probability of dying from cancer at 10 years from diagnosis for the most deprived male patients is approximately 36.5% (95% CI: 35.5%, 37.5%) for those aged 50 years old at diagnosis. This slightly increases to approximately 40.5% (95% CI: 39.8%, 41.1%) for those aged 80 years old at diagnosis. However, the largest change is in the probability of dying from other causes and CVD which have an increas-

ing contribution to the probability of dying from any cause for older male patients from the most (and least) deprived groups. For instance, the probability of dying from any cause by 10 years from diagnosis for the most deprived 50 year old male patients at diagnosis is 53.6% of which 17.1% is due to other causes and CVD. In contrast, the all-cause probability of death for the most deprived male patients aged 80 years old diagnosis is much higher at 92.5%. However, although the probability of dying due to cancer has only increased from 36.5% to 42.5%, much of the overall probability of dying is due to other causes (38.4%) and CVD (13.6%).

Absolute CIF differences between the most and least deprived male patients aged 50, 65 and 80 years old at diagnosis are presented on the third row of Figure 2. This shows that, for 50 year olds, the difference between CIFs for the most and least deprived groups are similar for deaths due to cancer and other causes. There is very little difference between the two deprivation groups for deaths due to CVD, however, this is due to a generally very low probability of death due to CVD. On the other hand, for older male patients, the difference in the probability of dying from other causes and CVD between the most and least deprived is larger and increases over time. This leads to a greater disparity in the probability of dying from other causes and CVD between the most and least deprived patients compared to the difference in the probability of dying due to cancer. Furthermore, after approximately 1 year from diagnosis for 65 year olds, and 2 years for 80 year olds, the difference in the probability of dying due to cancer for the most deprived compared to the least deprived patients reduces. This change in difference between the most and least deprived is greatest for the 80 year old male patients with cancer-specific CIF difference reducing from approximately 4.6% (95% CI: 4.2%, 5.0%) at 1 year from diagnosis to 3.2% (95% CI: 2.6%, 3.7%) by 10 years from diagnosis.

### *3.3.2 Restricted mean failure time and expected number of life-years lost due to a particular cause of death*

As discussed in Section 6, as a useful summary measure, the RMST estimate can be obtained. This is equivalent to the white area of the associated stacked plot in Figure 2 up to $t^*$ for a particular covariate pattern. Conversely, the area of the stacked areas give an estimate of the RMFT. The area of each of the partitioned stacks for each of the respective causes of death yield the expected life years lost due to cancer, CVD and other causes. These are presented for the most and least deprived 50, 65 and 80 year old male patients in Figure 3. Each of the stacks represent the average life-years lived in total and life-years lost due to a specific cause. The plots here present life-years lost and lived before different points in time up to 10 years from diagnosis. However, particular interest here is in the life-years lived, or lost, *before* 10 years from diagnosis. For example, total average life-years lived before 10 years from diagnosis for the most deprived 50 year old male patients is 3.99 years (95% CI: 3.84 years, 4.14 years). Of the 6.01 years of the total life-years lost, 2.72 years (95% CI: 2.60 years, 2.85 years) are due to cancer, 0.07 years (95% CI: 0.06 years, 0.09 years) are due to CVD and 1.19 (95% CI: 1.11 years, 1.28 years) due to other causes.

Table 1 presents differences in life-years lost due to each cause of death before 10 years from diagnosis between the most and least deprived groups for 50, 65 and 80 year olds, along with their associated 95% CIs. The absolute estimates of expected life-years lost for the most and least deprived patients at the individual ages are also presented. This provides us with an understanding of how many additional life-years most deprived patients are expected to lose due to a specific cause of death in comparison to the least deprived patients. For instance, at 10 years from diagnosis, 50 year old male patients from the most deprived group lose an additional 0.32 (95% CI: 0.28, 0.36) life-years due to cancer, 0.01 (95% CI: 0.01, 0.02) life-years

due to CVD and 0.33 (95% CI: 0.30, 0.36) life-years due to CVD compared to the least deprived group. For older male patients aged 80 years old, there is a greater disparity in life-years lost due to CVD (0.16 life-years) and other causes (0.76 life-years) between the most and least deprived.

## 3.4 Analysis of data with marginal estimates

When interest is in the covariate effects of particular groups, for example, between deprivation groups, it is useful to obtain standardised estimates as described in Section 2.4.4. By marginalising over the same covariate distribution, fairer comparisons can be made between particular covariate groups of interest. In this example, we standardise by age and sex in order to summarise the differences in survival between patients from the most and least deprived groups.

### *3.4.1 Cause-specific probability of death for the most deprived compared to the least deprived*

Figure 4 illustrates standardised CIFs stacked for each cause of death and Figure 5 presents absolute risk differences for each cause between the least and most deprived patients. As illustrated in Figure 4, patients from the most deprived group have a higher probability of dying from any cause (73.8%) compared to those from the least deprived group (63.3%). However, when partitioned into the different causes of death, the difference in total mortality between the most and least deprived groups is mostly due to other causes and CVD as indicated by the area proportions. The cause-specific marginal risk difference between the most and least deprived are presented in Figure 5 along with their respective 95% CIs. As can be seen here, the largest difference in risk is due to other causes and the largest difference in risk between the least and most deprived groups is due to other causes at 10 years from diagnosis (6.3%; 95% CI: 5.8%, 6.9%). Generally, the disparity in the probability of dying from other causes or CVD between the most and least deprived patients continues to increase over follow-up time. However, the cancer-specific risk difference

between the most and least deprived increases only for the first 2 years. After this point, the disparity in the probability of dying due to cancer between the most and least deprived begins to decrease.

### *3.4.2 Expected number of life-years lost for the most deprived compared to the least deprived*

In Figure 3, the expected life-years lost and total average life-years lived were presented for each cause of death before various time-points, $t^*$. By obtaining marginal estimates through standardisation over age and sex, we can focus on specific comparisons between the least and most deprived patients. The marginal expected life-years lived for each cause of death and total average life-years lived before each time, $t^*$, are similarly illustrated in Figure 6. If $t^* = 10$, then we have that the total average life-years lived before 10 years from diagnosis for the most deprived patients is 4.39 (95% CI: 3.78, 5.00). Of the 5.61 total expected life-years lost, 3.03 (95% CI: 2.66, 3.46) years are lost due to cancer, 0.46 (95% CI: 0.27, 0.81) years due to CVD and 2.11 (95% CI: 1.76, 2.53) years due to other causes. By obtaining marginal estimates of expected life-years lost, we are able to directly compare both deprivation groups and determine the additional life-years lost for patients that are the most deprived standardised by age and sex. Thus, where $t^* = 10$, we have that the additional life-years lost due to cancer, CVD and other causes before 10 years from diagnosis for the most deprived patients is 0.31 (95% CI: 0.25, 0.37), 0.05 (95% CI: 0.02, 0.08) and 0.44 (95% CI: 0.33, 0.54) life-years respectively.

## 4 Discussion

This paper presents novel estimation of RMLT and expected life-years lost from within the flexible parametric survival modelling framework in the presence of competing risks. This can be done either on the CSHs or cumulative incidence scale and allows easy incorporation of time-dependent effects to relax the proportionality assumption. These also offer additional advantages over the more popular Cox PH

and Fine and Gray models [7, 5]. In particular, we illustrate how one can easily obtain comparative predictions based on the expected number of life-years lost due to a specific cause of death in addition to other useful estimands, such as absolute differences in the cumulative incidence functions. A common approach for obtaining marginal estimates uses inverse probability weighted estimating equations. However, different estimators need to be calculated subject to whether it is of interest to obtain marginal or non-marginal/conditional estimates [48, 49]. On the other hand, marginal estimates using standardisation are easily obtained in addition to conditional estimates within the FPM approach from a single model. FPMs in both a standard survival analysis and for competing risks data offer numerous advantages in prediction, specifically, through its estimation of the baseline hazard function using RCS and easy inclusion of time-dependent effects. In spite of this, it is also important to consider limitations that are often highlighted. One such limitation is the problem of choosing the appropriate number of knots for the underlying baseline hazard function using RCS, and for when including time-dependent effects when relaxing the proportional hazards assumption. However, a number of extensive simulation studies have been carried out evaluating how many knots are required in order to accurately capture (both simple and complex i.e. time-dependent) shapes of the baseline hazard function. For instance, Bower et. al. [37] and Syriopoulou et. al. [50] both conclude predictions are not sensitive to the choice in the number of knots, provided that a sufficient number of degrees of freedom are used. In other words, too few degrees of freedom may be too simple to accurately capture the effect, and too many will lead to over-fitting. As a guideline, 5 degrees of freedom to capture baseline effects and 3 degrees of freedom for any time-dependent effects are suggested as a starting point. However, it is further suggested that for each individual study, sensitivity analyses are carried out in order to assess model fit and robustness to the choice in degrees of freedom [37, 50]. Syriopoulou et. al

[50] also reach similar conclusions with extension to marginal model-based estimates when obtaining predictions using standardisation. Alternatively, a penalised approach for choosing the appropriate number of degrees of freedom for RCS can be used [51]. The interpretation of the RMLT measure also has some notable limitations. Although communication in terms of changes in life-years lost to clinicians and patients rather than probabilities is attractive, applying an upper bound, $t^*$, to the time interval may add some difficulty in understanding of the measure. This is because, RMLT for an arbitrary choice of $t^*$ can only be used to estimate the average risk within a restricted time period for a group of patients. Furthermore, it should be highlighted that the expected life-years lost makes comparison with an immortal cohort where patients are alive for the whole interval from 0 to time $t^*$. A similar "unrestricted" measure that do not compare to an immortal cohort can be estimated within the relative survival framework based on extrapolation of the excess hazard rate. This is usually referred to as the number of life years lost, or the loss in expectation of life and is calculated based on a comparison of the life-expectancy of cancer patients to a comparable population group who are assumed to be cancer-free [52, 53, 54]. However, this relies on the assumption that this extrapolation is appropriate which is not made for the RMLT estimate. In addition to the above, due to the dependence of the interpretation of RMST on follow-up time, comparison between different studies, for example, between countries, becomes difficult. It has also been further shown that the difference in RMST between two covariate groups depends on the outcome rates within each group. Therefore, it is recommended that differences in RMST, RMFT and expected number of life-years lost, are reported alongside their respective survival, or cumulative incidence functions, in order to allow comparability and to obtain the entire picture of the impact of different groups on outcome [24]. This further points to additional advantages of

estimation of RMFT within the flexible parametric modelling framework, as these additional measures are easily obtained from the same model.

## 5 Conclusions

The RMLT measure is presented as a useful summary measure with an attractive interpretation which can aid in the analysis of competing risks data. As discussed by others, it is also useful to present estimated cause-specific CIFs alongside CSHs [6, 34]. We propose FPMs as the chosen estimator as it allows easy estimation of various estimands from a single model providing both conditional and marginal estimates. Note that, although not discussed here, if appropriate confounders are adjusted for, one can also infer causal effects between two groups using standardisation. However, one must also consider the additional complexities and issues in interpretation with the inclusion of time-dependent risk-groups [44]. Furthermore, the RMLT measure can be easily extended for obtaining conditional estimates, for example, the average life-years lived before $t^*$ years given survival to time $t_0$ from diagnosis. Example Stata code for the model and prediction of measures provided in this paper is outlined in Appendix D.

## 6 List of Abbreviations

| | |
|---|---|
| CSH | Cause-specific hazards |
| CIF | Cumulative incidence function |
| CI | Confidence interval |
| CVD | Cardiovascular disease |
| DF | Degrees of freedom |
| FPM | Flexible parametric model |
| FPCIM | Flexible parametric cumulative incidence model |
| HR | Hazard ratio |
| LYL | Life-years lost |
| RCS | Restricted cubic splines |
| RMFT | Restricted mean failure time |
| RMLT | Restricted mean lifetime |
| RMST | Restricted mean survival time |
| SDH | Subdistribution hazards |
| SHR | Subdistribution hazard ratio |

**Declarations**

Ethics approval and consent to participate

The study received ethical approval from North West - Greater Manchester East Research Ethics Committee (14/NW/1449).

Consent for publication

Not applicable.

Availability of data and material

The data that support the findings of this study are available from Public Health England (https://www.gov.uk/government/publications/accessing-public-health-england-data/about-the-phe-odr-and-accessing-data), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Competing interests

SIM works at Roche 0.5 WTE (working-time-equivalent). All other authors declare that they have no competing interests.

Funding

This project was funded by Cancer Research UK [Grant Number C1483/A18262] and supported the research presented in this article. Cancer Research UK had no role in the methods development, study design, data collection, analysis and interpretation of the data, or in the writing of the article. The article processing charge for open access was funded by Cancer Research UK as a member of the 2020/21 Charity Open Access Fund (COAF).

Author's contributions

Contributions of each author to the work detailed in the manuscript are as follows. All authors, SIM, MJR and PCL, equally contributed towards the conception and design, acquisition of data, and interpretation of data. Specifically, SIM carried out the analysis of the data and prepared the manuscript. PCL and MJR supervised the project and contributed comments and suggestions throughout. All authors were involved in the drafting and submission of this research article.

Acknowledgements

Not applicable.

**Author details**

[1]Biostatistics Research Group, Department of Health Sciences, University of Leicester, University Road, LE1 7RH Leicester, UK. [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

**References**

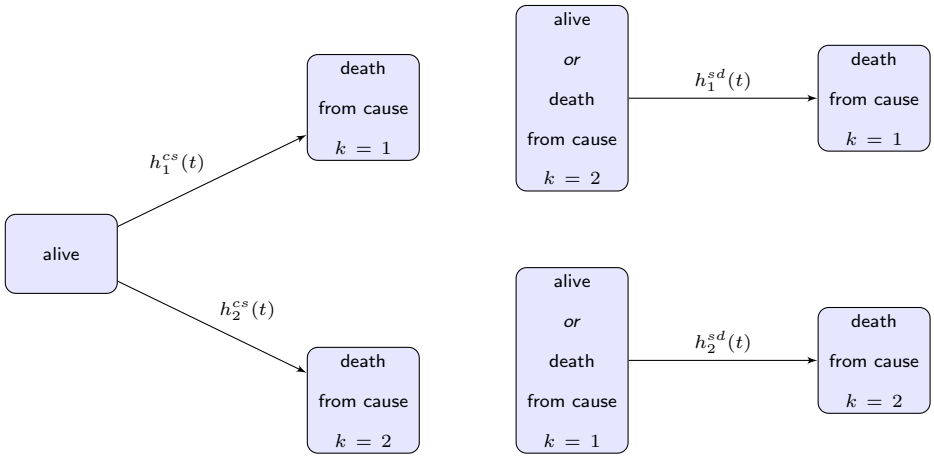
1. Putter, H., Fiocco, M., Geskus, R.B.: Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine **26**(11), 2389–2430 (2007). doi:10.1002/sim.2712
2. Geskus, R.: Data Analysis with Competing Risks and Intermediate States. CRC Press, Boca Raton (2016)
3. Andersen, P.K.: Decomposition of number of life years lost according to causes of death. Statistics in Medicine **32**, 5278–85 (2013). doi:10.1002/sim.5903
4. Wolbers, M., Koller, M.T., Stel, V.S., Schaer, B., Jager, K.J., Leffondré, K., Heinze, G.: Competing risks analyses: objectives and approaches. European Heart Journal **35**(42), 2936–2941 (2014). doi:10.1093/eurheartj/ehu131
5. Mozumder, S.I., Rutherford, M., Lambert, P.: Direct likelihood inference on the cause-specific cumulative incidence function: A flexible parametric regression modelling approach. Statistics in Medicine **37**(1), 82–97 (2017). doi:10.1002/sim.7498
6. Lambert, P.C., Wilkes, S.R., Crowther, M.J.: Flexible parametric modelling of the cause-specific cumulative incidence function. Statistics in Medicine **36**(9), 1429–1446 (2016). doi:10.1002/sim.7208
7. Hinchliffe, S.R., Lambert, P.C.: Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. BMC Medical Research Methodology **13**(1) (2013). doi:10.1186/1471-2288-13-13
8. Putter, H., Sasako, M., Hartgrink, H.H., van de Velde, C.J.H., van Houwelingen, J.C.: Long-term survival with non-proportional hazards: results from the Dutch gastric cancer trial. Stat Med **24**(18), 2807–2821 (2005). doi:10.1002/sim.2143
9. Royston, P., Parmar, M.K.B.: Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine **21**(15), 2175–2197 (2002). doi:10.1002/sim.1203
10. Sauerbrei, W., Royston, P., Look, M.: A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. Biometrical Journal **49**(3), 453–473 (2007). doi:10.1002/bimj.200610328. https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.200610328

11. Hernán, M.A.: The hazards of hazard ratios. Epidemiology **21**(1), 13–15 (2010). doi:10.1097/EDE.0b013e3181c1ea43
12. Belot, A., Ndiaye, A., Luque-Fernandez, M.-A., Kipourou, D.-K., Maringe, C., Rubio, F.J., Rachet, B.: Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. Clinical Epidemiology **Volume 11**, 53–65 (2019). doi:10.2147/clep.s173523
13. Calkins, K.L., Canan, C.E., Moore, R.D., Lesko, C.R., Lau, B.: An application of restricted mean survival time in a competing risks setting: comparing time to ART initiation by injection drug use. BMC Medical Research Methodology **18**(1) (2018). doi:10.1186/s12874-018-0484-z
14. Karrison, T.: Restricted mean life with adjustment for covariates. Journal of the American Statistical Association **82**(400), 1169–1176 (1987). doi:10.1080/01621459.1987.10478555. https://www.tandfonline.com/doi/pdf/10.1080/01621459.1987.10478555
15. Royston, P., Parmar, M.K.B.: The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Satistics in Medicine **30**(19), 2409–2421 (2011). doi:10.1002/sim.4274
16. Royston, P., Parmar, M.K.B.: Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Medical Research Methodology **13**, 152 (2013). doi:10.1186/1471-2288-13-152
17. Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M., Wei, L.-J.: Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. Journal of Clinical Oncology **32**(22), 2380–2385 (2014). doi:10.1200/JCO.2014.55.2208
18. Chen, P.-Y., Tsiatis, A.A.: Causal inference on the difference of the restricted mean lifetime between two groups. Biometrics **57**(4), 1030–1038 (2001). doi:10.1111/j.0006-341x.2001.01030.x
19. Dehbi, H.-M., Royston, P., Hackshaw, A.: Life expectancy difference and life expectancy ratio: two measures of treatment effects in randomised trials with non-proportional hazards. BMJ, 2250 (2017). doi:10.1136/bmj.j2250
20. Austin, P.C.: The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. Statistics in Medicine **33**(7), 1242–1258 (2014). doi:10.1002/sim.5984. https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.5984
21. Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research **46**(3), 399–424 (2011). doi:10.1080/00273171.2011.568786. PMID: 21818162. https://doi.org/10.1080/00273171.2011.568786
22. Beltrán-Sánchez, H., Preston, S.H., Canudas-Romo, V.: An integrated approach to cause-of-death analysis: cause-deleted life tables and decompositions of life expectancy. Demographic Research **19**, 1323–1350 (2008). doi:10.4054/demres.2008.19.35
23. Andersen, P.K., Perme, M.P.: Pseudo-observations in survival analysis. Statistical Methods in Medical Research **19**(1), 71–99 (2010). doi:10.1177/0962280209105020
24. Kloecker, D.E., Davies, M.J., Khunti, K., Zaccardi, F.: Uses and limitations of the restricted mean survival time: Illustrative examples from cardiovascular outcomes and mortality trials in type 2 diabetes. Annals of Internal Medicine **172**(8), 541 (2020). doi:10.7326/m19-3286
25. Royston, P.: Flexible Parametric Survival Analysis Using Stata : Beyond the Cox Model. Stata Press, College Station, TX (2011)

26. Lambert, P.C., Dickman, P.W., Nelson, C.P., Royston, P.: Estimating the crude probability of death due to cancer and other causes using relative survival models. Stat Med **29**, 885–895 (2010)
27. Rothman, K.: Modern Epidemiology. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia (2008)
28. Sjölander, A.: Regression standardization with the r package stdReg. European Journal of Epidemiology **31**(6), 563–574 (2016). doi:10.1007/s10654-016-0157-3
29. Gray, R.J.: A class of k-sample tests for comparing the cumulative incidence of a competing risk. The Annals of Statistics **16**, 1141–1154 (1988)
30. Andersen, P.K., Keiding, N.: Interpretability and importance of functionals in competing risks and multistate models. Statistics in Medicine **31**(11-12), 1074–1088 (2012). doi:10.1002/sim.4385
31. Beyersmann, J., Allignol, A., Schumacher, M.: Competing Risks and Multistate Models with R. Springer, New York (2012). doi:10.1007/978-1-4614-2035-4. https://doi.org/10.1007/978-1-4614-2035-4
32. Andersen, P.K., Geskus, R.B., de Witte, T., Putter, H.: Competing risks in epidemiology: possibilities and pitfalls. International Journal of Epidemiology **41**(3), 861–870 (2012). doi:10.1093/ije/dyr213
33. Austin, P.C., Fine, J.P.: Practical recommendations for reporting fine-gray model analyses for competing risk data. Statistics in Medicine **36**(27), 4391–4400 (2017). doi:10.1002/sim.7501
34. Latouche, A., Allignol, A., Beyersmann, J., Labopin, M., Fine, J.P.: A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. Journal of Clinical Epidemiology **66**(6), 648–653 (2013). doi:10.1016/j.jclinepi.2012.09.017
35. Zhao, L., Claggett, B., Tian, L., Uno, H., Pfeffer, M.A., Solomon, S.D., Trippa, L., Wei, L.J.: On the restricted mean survival time curve in survival analysis. Biometrics **72**(1), 215–221 (2015). doi:10.1111/biom.12384
36. Rutherford, M.J., Abel, G.A., Greenberg, D.C., Lambert, P.C., Lyratzopoulos, G.: The impact of eliminating age inequalities in stage at diagnosis on breast cancer survival for older women. British Journal of Cancer **112 Suppl**, 124–128 (2015). doi:10.1038/bjc.2015.51
37. Bower, H., Crowther, M.J., Rutherford, M.J., Andersson, T.M.-L., Clements, M., Liu, X.-R., Dickman, P.W., Lambert, P.C.: Capturing simple and complex time-dependent effects using flexible parametric survival models: A simulation study. Communications in Statistics - Simulation and Computation **0**(0), 1–17 (2019). doi:10.1080/03610918.2019.1634201. https://doi.org/10.1080/03610918.2019.1634201
38. Lambert, P.C., Royston, P.: Further development of flexible parametric models for survival analysis. The Stata Journal: Promoting communications on statistics and Stata **9**(2), 265–290 (2009). doi:10.1177/1536867x0900900206
39. Crowther, M.J., Lambert, P.C.: A general framework for parametric survival analysis. Stat Med **33**(30), 5280–5297 (2014). doi:10.1002/sim.6300
40. Young, J.G., Stensrud, M.J., Tchetgen, E.J.T., Hernn, M.A.: A causal framework for classical statistical estimands in failure time settings with competing events (2018). 1806.06136
41. Stensrud, M.J., Aalen, J.M., Aalen, O.O., Valberg, M.: Limitations of hazard ratios in clinical trials. European Heart Journal **40**(17), 1378–1383 (2018). doi:10.1093/eurheartj/ehy770
42. Valeri, L., Chen, J.T., Garcia-Albeniz, X., Krieger, N., VanderWeele, T.J., Coull, B.A.: The role of stage at diagnosis in colorectal cancer black-white survival disparities: A counterfactual causal inference approach. Cancer Epidemiology Biomarkers & Prevention **25**(1), 83–89 (2015). doi:10.1158/1055-9965.epi-15-0456
43. Gran, J.M., Lie, S.A., Ãyeflaten, I., Borgan, Ã., Aalen, O.O.: Causal inference in multi-state models-sickness absence and work for 1145 participants after work rehabilitation. BMC Public Health **15**, 1082 (2015). doi:10.1186/s12889-015-2408-8

44. von Cube, M., Schumacher, M., Wolkewitz, M.: Causal inference with multistate models—estimands and estimators of the population attributable fraction. Journal of the Royal Statistical Society: Series A (Statistics in Society) (2019). doi:10.1111/rssa.12486
45. Geskus, R.B.: Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. Biometrics **67**(1), 39–49 (2011). doi:10.1111/j.1541-0420.2010.01420.x
46. Jeong, J.-H., Fine, J.P.: Parametric regression on cumulative incidence function. Biostatistics **8**(2), 184–196 (2007). doi:10.1093/biostatistics/kxj040
47. Mozumder, S.I., Rutherford, M.J., Lambert, P.C.: A flexible parametric competing-risks model using a direct likelihood approach for the cause-specific cumulative incidence function. The Stata Journal: Promoting communications on statistics and Stata **17**(2), 462–489 (2017). doi:10.1177/1536867x1701700212
48. Schaubel, D.E., Wei, G.: Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring. Biometrics **67**(1), 29–38 (2010). doi:10.1111/j.1541-0420.2010.01449.x
49. Zhang, M., Schaubel, D.E.: Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. Biometrics **68**(4), 999–1009 (2012)
50. Syriopoulou, E., Mozumder, S.I., Rutherford, M.J., Lambert, P.C.: Robustness of individual and marginal model-based estimates: A sensitivity analysis of flexible parametric models. Cancer Epidemiology **58**, 17–24 (2019). doi:10.1016/j.canep.2018.10.017
51. Clements, M., Liu, X.-R.: Rstpm2: Smooth Survival Models, Including Generalized Survival Models. (2019). R package version 1.5.1. https://CRAN.R-project.org/package=rstpm2
52. Andersson, T.M.-L., Dickman, P.W., Eloranta, S., Lambe, M., Lambert, P.C.: Estimating the loss in expectation of life due to cancer using flexible parametric survival models. Statistics in Medicine **32**(30), 5286–5300 (2013). doi:10.1002/sim.5943
53. Burnet, N.G., Jefferies, S.J., Benson, R.J., Hunt, D.P., Treasure, F.P.: Years of life lost (YLL) from cancer is an important measure of population burden – and should be considered when allocating research funds. British Journal of Cancer **92**(2), 241–245 (2005). doi:10.1038/sj.bjc.6602321
54. Chu, P.-C., Wang, J.-D., Hwang, J.-S., Chang, Y.-Y.: Estimation of life expectancy and the expected years of life lost in patients with major cancers: extrapolation of survival curves under high-censored rates. Value Health **11**(7), 1102–1109 (2008). doi:10.1111/j.1524-4733.2008.00350.x
55. Abramowitz, M.: Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables. Dover Publications, New York (1965)
56. Lambert, P.C.: The estimation and modeling of cause-specific cumulative incidence functions using time-dependent weights. The Stata Journal: Promoting communications on statistics and Stata **17**(1), 181–207 (2017). doi:10.1177/1536867x1701700110
57. Lunn, M., McNeil, D.: Applying Cox regression to competing risks. Biometrics **51**(2), 524–532 (1995)


**Figures**




(a) Transitions described by the cause-specific hazard rate. The typical competing risks scenario.

(b) Transitions described by the subdistribution hazard rate.

Figure 1: Schematics of transitions from an initial state to one of $K = 2$ causes of death.

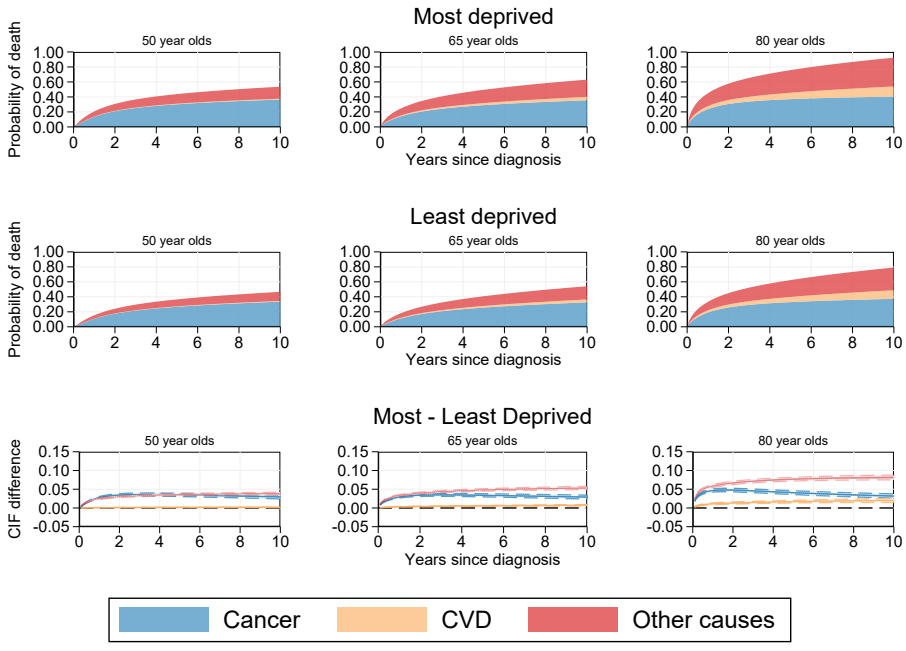



Figure 2: Stacked cause-specific CIFs by deprivation group and CIF differences for male patients at specific ages.

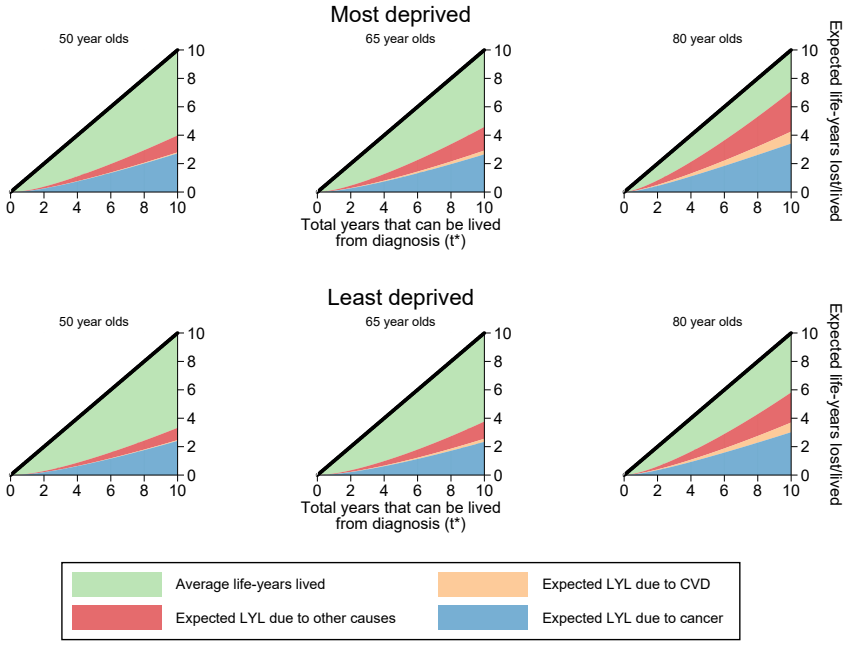



Figure 3: Stacked plots of expected life-years lost partitioned by each cause of death for male patients.

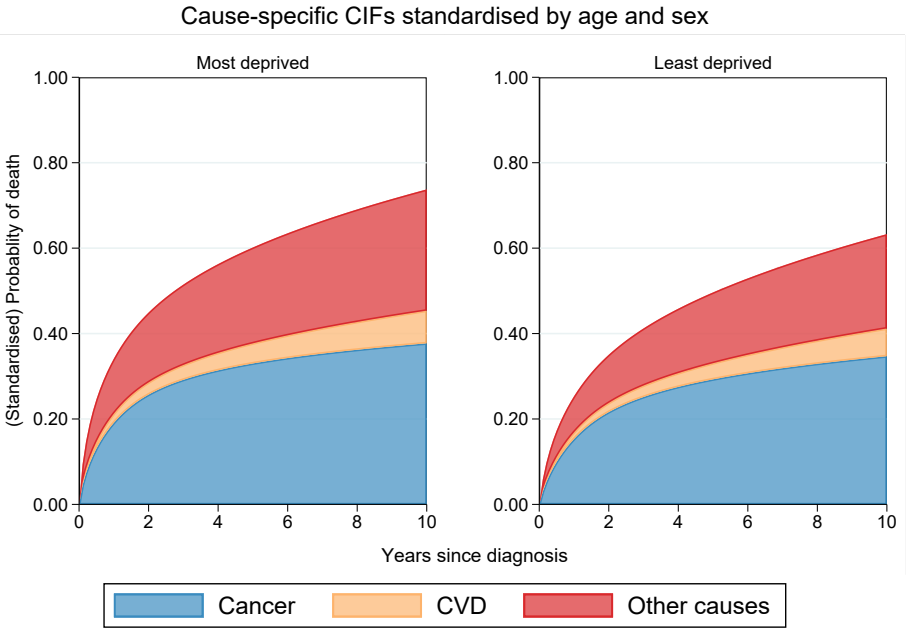



Figure 4: Estimated cause-specific CIFs standardised by age and sex for each deprivation group.

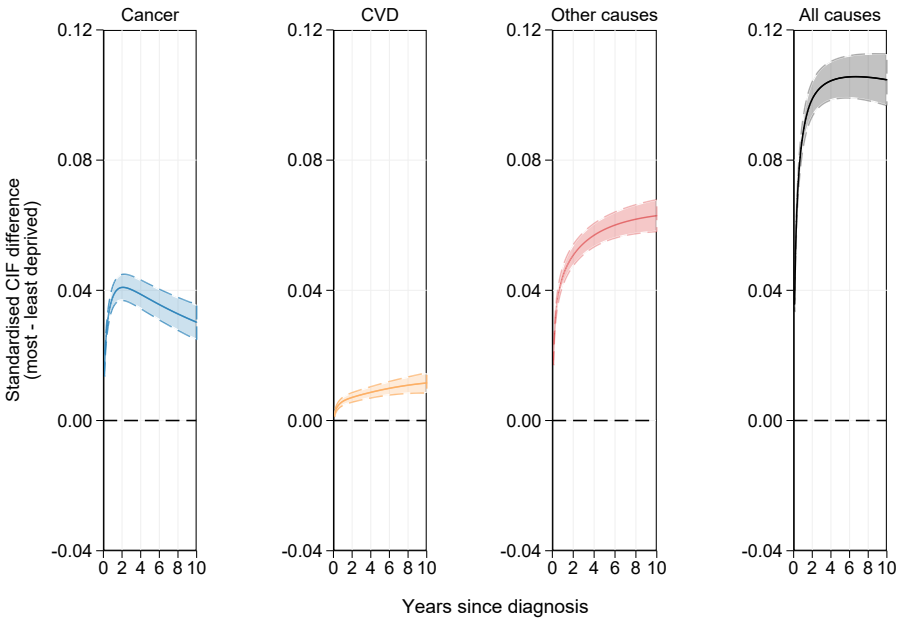



Figure 5: Estimated CIF differences for each cause of death standardised by age and sex with 95% CIs.

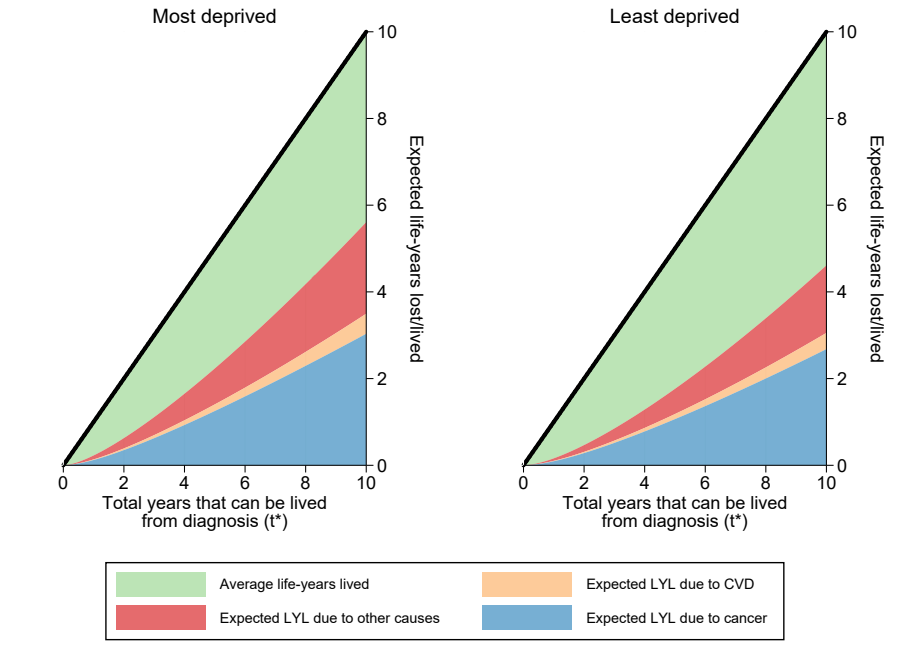



Figure 6: Stacked plots of expected life-years lost for each cause of death standardised by age and sex.

718 **Tables**

Table 1: Expected LYL for each cause for males aged 50, 65 and 80 years old at diagnosis.

| | Most Deprived | | | Least Deprived | | | Difference | | |
|---|---|---|---|---|---|---|---|---|---|
| | **LYL** | **95% LCI** | **95% UCI** | **LYL** | **95% LCI** | **95% UCI** | **LYL** | **95% LCI** | **95% UCI** |
| **50 Yrs Old** | | | | | | | | | |
| Cancer | 2.724 | [2.604, | 2.848] | 2.407 | [2.299, | 2.519] | 0.317 | [0.277, | 0.357] |
| CVD | 0.069 | [0.055, | 0.088] | 0.056 | [0.044, | 0.071] | 0.014 | [0.009, | 0.018] |
| Other causes | 1.195 | [1.113, | 1.282] | 0.864 | [0.804, | 0.929] | 0.330 | [0.300, | 0.361] |
| **65 Yrs Old** | | | | | | | | | |
| Cancer | 2.654 | [2.179, | 3.232] | 2.340 | [1.913, | 2.864] | 0.313 | [0.250, | 0.377] |
| CVD | 0.271 | [0.149, | 0.495] | 0.219 | [0.120, | 0.400] | 0.052 | [0.019, | 0.085] |
| Other causes | 1.662 | [1.285, | 2.149] | 1.212 | [0.930, | 1.580] | 0.449 | [0.339, | 0.559] |
| **80 Yrs Old** | | | | | | | | | |
| Cancer | 3.415 | [3.055, | 3.818] | 3.018 | [2.690, | 3.386] | 0.397 | [0.340, | 0.454] |
| CVD | 0.840 | [0.468, | 1.508] | 0.681 | [0.378, | 1.228] | 0.159 | [0.063, | 0.255] |
| Other causes | 2.845 | [2.426, | 3.337] | 2.120 | [1.792, | 2.508] | 0.725 | [0.618, | 0.833] |

## Appendix A: Restricted cubic spline variables

Given a vector of $M$ knots, $\mathbf{m}$ and a vector of $M-1$ parameters, $\boldsymbol{\gamma}$, with $M-1$ degrees of freedom (df), the restricted cubic spline function, $s(\ln(t);\boldsymbol{\gamma},\mathbf{m})$, is defined as,

$$s(\ln(t);\boldsymbol{\gamma},\mathbf{m}) = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_{(M-1)} z_{(M-1)} \tag{31}$$

Where $z_1, \cdots, z_{(M-1)}$ are the basis functions of the restricted cubic splines and are defined as,

$$z_1 = \ln(t) \tag{32}$$

$$z_j = (\ln(t) - m_j)_+^3 - \phi_j(\ln(t) - m_1)_+^3 - (1-\phi_j)(\ln(t) - m_M)_+^3, \quad j = 2, \cdots, M-1$$

where,

$$\phi_j = \frac{m_M - m_j}{m_M - m_1} \tag{33}$$

and

$$(u)_+ = \begin{cases} u, & \text{if } u > 0 \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

Usually, $M$ knots are placed at equally spaced centiles of the distribution of the uncensored log-survival times including two boundary knots at the $0^{th}$ and $100^{th}$ centiles.

## Appendix B: Gaussian quadrature

With the general Gaussian quadrature rule, the integral of any polynomial function, $g(u)$, over the interval $[-1, 1]$ can be evaluated. This performs best for integrals that can be approximated by a polynomial function of degree $2m-1$, where $m$ is a pre-determined number of points, otherwise known as nodes, or abscissae. Hence, this integral can be evaluated for,

$$\int_{-1}^{1} g(u)\mathrm{d}u = \int_{-1}^{1} W(u)g(u)\mathrm{d}u \tag{35}$$

where, $W(u)$, is a known weighting function. Here, the integral, e.g. the cause-specific cumulative incidence function, is calculated using Gauss-Legendre quadrature, with $W(u) = 1$. With this, based on a set of pre-defined number of nodes, $u_i'$, and associated Lagrange polynomials of degree $m$, $P_m(u)$, weights, $w_i'$, for $i = 1, \ldots, m$, are obtained such that,

$$w_i^{'} = \frac{2}{\left(1 - u_i^{'2}\right)\left(P_m'(u_i')\right)^2} \tag{36}$$

and are provided by Abramowitz and Stegun [55]. Therefore, equation 35 is approximated by,

$$\int_{-1}^{1} g(u)\mathrm{d}u \approx \sum_{i=1}^{m} w_i' g(u_i') \tag{37}$$

However, for survival data, functions are evaluated over an interval $[0, t]$. Therefore, to apply the Gaussian quadrature rule in equation 35, integrals over the interval $[0, t]$ must be changed to an interval over $[-1, 1]$ such that,

Table 2: Distribution of data on key covariates included in the analysis for n = 159,022 patients

| | Females, n(%) | Least deprived, n(%) | Age, mean(sd) |
|---|---|---|---|
| Cancer | 21 137 (43.27) | 25 084 (51.35) | 72.25 (10.57) |
| CVD | 3 158 (39.54) | 3 853 (48.24) | 76.78 (7.96) |
| Other Causes | 13 716 (42.71) | 14 955 (46.57) | 74.04 (9.64) |
| All Causes | 38 011 (42.74) | 43 892 (49.35) | 73.30 (10.13) |
| Alive/Censored within 10 yrs | 30 663 (43.76) | 43 079 (61.47) | 68.05 (9.97) |
| Total | 68 974 (43.19) | 86 971 (54.59) | 70.99 (10.39) |

$$\int_0^t g(u)\mathrm{d}u = \frac{t-0}{2}\int_{-1}^{1} g\left(\frac{t-0}{2}u + \frac{t+0}{2}\right)\mathrm{d}u \tag{38}$$

Therefore, a function evaluated at $t_1, \ldots, t$ different time-points over an interval $[0, t]$ is approximated by applying Gaussian quadrature rules with $W(u) = 1$ such that,

$$\int_0^{t_\omega} g(u)\mathrm{d}u \approx \frac{t-0}{2}\sum_{i=1}^{m} w_i' g\left(\frac{t-0}{2}u_i' + \frac{t+0}{2}\right) \tag{39}$$

## Appendix C: Additional summary statistics

Table 2 provide summary statistics on the distribution of key covariates of interest for inclusion in analysis i.e. sex, deprivation group (least/most deprived) and age, by cause of death, and in total.

Figure 7 represents the cause-specific cumulative incidence functions estimates obtained by the non-parametric Aalen-Johansen estimator. This summarises the probability of dying from each cause of death by sex and deprivation groups.

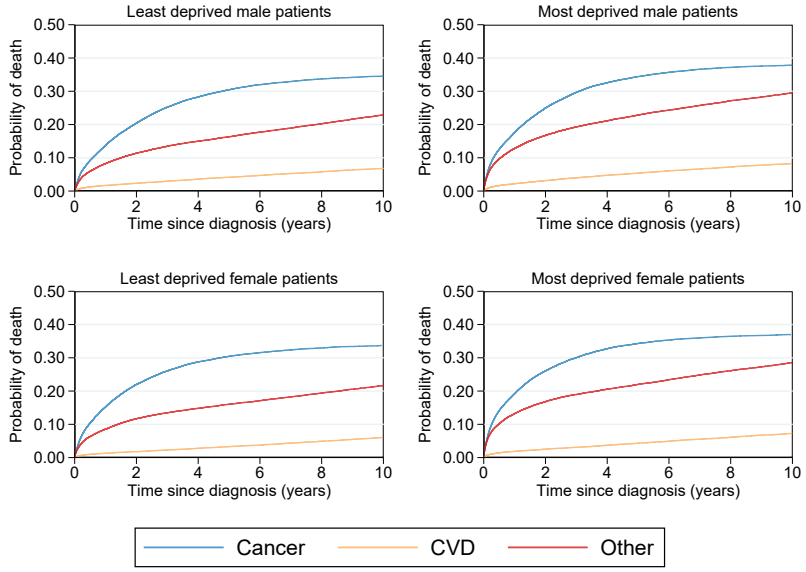



Figure 7: Cause-specific cumulative incidence functions (CIFs) Aalen-Johansen estimates for each cause of death.

Figure 8 illustrates the all-cause survival probabilities obtained by the non-parametric Kaplan-Meier estimator. This summarises the all-cause probability of survival by sex and deprivation groups.

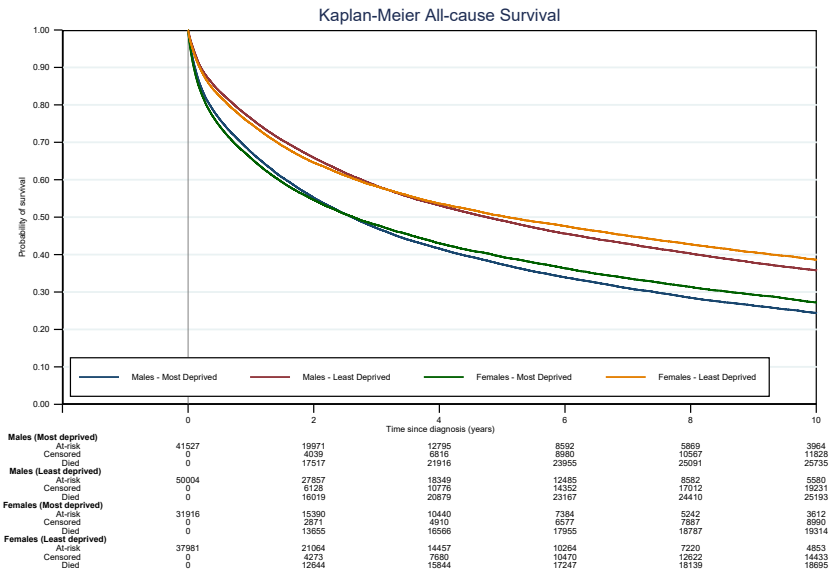



Figure 8: All-cause Kaplan-Meier survival probabilities by sex and deprivation groups.

## Appendix D: Stata code for obtaining predictions

This appendix outlines Stata code used to obtain predictions presented in the paper. Some user-defined Stata commands are required which can be installed from the Boston College Statistical Software Components (SSC) archive by calling,

`ssc install` [*command*]

The following must be installed before running the code:

- `stpm2`: To fit the flexible parametric models described in Section 2.3.
- `rcsgen`: To generate the restricted cubic spline functions.
- `stcrprep`: To restructure data and calculate time-dependent censoring weights in order to fit models on the subdistribution hazards scale using standard Stata commands.

To obtain marginal (and non-marginal) estimates using standardisation, the `standsurv` command must be installed. This will be released on SSC soon, however, in the meantime, it can be installed by running,

```
net from https://www.pclambert.net/downloads/standsurv
```

### *D.0.1 Preparing the data for analysis*

To prepare the data for a survival analysis in Stata, we must first run the `stset` command. We identify the variable that records survival time (in days), `exit2`, the indicator variable for cause of death, `cod`, where death from `cancer = 1`, `CVD = 2` and `other causes = 3` and finally the variable for date of diagnosis, `dx`. The `scale` option is used to transform the survival time into years from days and we use the `exit` option to restrict follow-up time to 10 years from diagnosis and censor those still alive at 2014. In order to ensure that the death indicator, `_d`, generated after `stset` matches the death indicator for cause of death, we create a new cause of death indicator, `cod2`, so that those who die either after 10 years from diagnosis or 2014 are administratively censored. Finally, to generate restricted cubic spline

variables for the non-linear effect of age centred at 45 years old at diagnosis, we use `rcsgen`. For 3 degrees of freedom, 3 new age spline variables are created, `rcsage1` − `rcsage3`, and we store knot positions and matrix for orthogonalization which are required for post-estimation predictions at specific ages.

```
stset exit2, failure(cod=1,2,3) id(id) scale(365.25) origin(dx) ///
>   exit(time min(dx + 365.25*10.01,mdy(12,31,2013)))

//must ensure that those that die after follow-up time
//   are administratively censored
gen cod2 = cond(_d==0,0,cod)

//center non-linear age (rcsage) at 45 years old
rcsgen age, gen(rcsage) df(3) orthog center(45)
//store knot positions in global macro
global knots `r(knots)´
//save matrix for orthogonalization
matrix Rage = r(R)
```

To restructure the data and calculate the time-dependent censoring weights so that we may fit a model on the subdistribution hazards scale, we use `stcrprep`[56]. Here, we specify `wtstpm2` to estimate the censoring distribution using a Royston-Parmar flexible parametric model with covariates included in the `censcov` option. The data is restructured based on the variable `failcode`, which splits the data according to the cause of interest. This is used to fit identify for which cause the model is to be fitted for. For clarity, we create dummy variables for each of the causes of death from `failcode` and generate `_cancer`, `_cvd` and `_other`. Another indicator variable, `event`, is also created to identify at which split time interval, or row, death (from any cause) is observed for that patient. To incorporate the calculated weights from `stcrprep`, we must `stset` the data again with `tstart` and `tstop`. These are also provided by `stcrprep` and give the times at which an individual starts and stops being at risk.

```
stcrprep, events(cod2) keep(age mostdep sex rcsage?) trans(1 2 3) ///
>    wtstpm2 censcov(mostdep sex rcsage?) every(1)
```

```
gen event = cod2 == failcode

stset tstop [iw=weight_c], failure(event) enter(tstart) noshow

tab failcode, gen(cause)
rename cause1 _cancer
rename cause2 _cvd
rename cause3 _other
```

### *D.0.2 Model*

The model described in Section 2.3 can be fitted in two ways after preparing the data. We can either fit separate models for each of the causes of death, or fit a single model to cancer, CVD and other causes simultaneously. Here, we demonstrate for the latter to make illustration of the code for obtaining predictions post-estimation easier. However, in order to fit the equivalent single model with coefficients comparable to the models fitted individually to each of the causes of death, the knot locations on the cause-specific survival time distributions must be stored. These are stored in global macros for each of the causes of death.

```
global knotstvc_opt
global bknotstvc_opt

foreach cause in cancer other cvd {
 2.   global lnbhknots_`cause´
 3. }

foreach cause in cancer other cvd {
 2.   stpm2 mostdep sex rcsage? if _`cause´==1, df(3) ///
 >       tvc(mostdep sex rcsage?) dftvc(2) scale(h) eform
 3.   global bhknots_`cause´ `e(bhknots)´
 4.   global boundknots_`cause´ `e(boundary_knots)´
 5.   foreach cov in mostdep sex rcsage1 rcsage2 rcsage3 {
 6.      global knotstvc_opt ${knotstvc_opt} ///
 >          `cov´_`cause´ `e(tvcknots_`cov´)´
 7.   }
 8.   global knotstvc_opt ${knotstvc_opt} _`cause´ ${bhknots_`cause´}
 9.   global bknotstvc_opt ${bknotstvc_opt} _`cause´ ${boundknots_`cause´}
10. }
```

Here we define a global macro of the list of covariates to be included in the single model. As the data is stacked, interactions need to be created between the covariates and the indicator variable for each cause of death. See Lunn and McNeil[57] for further details. The baseline coefficient, i.e. the constant in the cause-specific model, is calculated in `_cancer`, `_cvd` and `_other`. We therefore fit a model for each of the causes of death simultaneously without a constant using `nocons` and the baseline splines using `rcsbaseoff`. Instead, the baseline splines are specified as time-dependent splines for the coefficient that corresponds to the constant in its respective model for that particular cause of death. These were stored in the global macro `bknotstvc_opt`. Since knots are specified according to the time scale, rather than the log-time scale, the `knscale(time)` option is used.

```
. global covlist
. global covlist_tvc

. foreach cause in cancer cvd other {
  2.   global covlist $covlist _`cause´
  3.   global covlist_tvc $covlist_tvc _`cause´
  4.   foreach cov in mostdep sex rcsage1 rcsage2 rcsage3 {
  5.      gen `cov´_`cause´ = `cov´*_`cause´
  6.      global covlist $covlist `cov´_`cause´
  7.      global covlist_tvc $covlist_tvc `cov´_`cause´
  8.   }
  9. }

. di "$covlist"
_cancer mostdep_cancer sex_cancer rcsage1_cancer rcsage2_cancer rcsage3_cancer
_cvd mostdep_cvd sex_cvd rcsage1_cvd rcsage2_cvd rcsage3_cvd
_other mostdep_other sex_other rcsage1_other rcsage2_other rcsage3_other

. stpm2 $covlist ///
>   , scale(h) tvc($covlist_tvc) knotstvc(${knotstvc_opt}) ///
>    bknotstvc(${bknotstvc_opt}) knscale(time) rcsbaseoff eform nocons
```

### *D.0.3 Predictions*

Although `standsurv` was written for obtaining marginalised predictions, it can also be used to obtain non-marginalised estimates. This is done by simply specifying the

entire covariate pattern so that the predictions are not averaged over any covariate distribution. To obtain predictions at a specific age, we need to calculate the spline variables at that particular age centred at 45 years old with the same knot locations and projection matrix as before. The spline variables are stored in the local macros `c1`, `c2` and `c3`. An example is given below when the cause of interest is cancer and we want to make comparisons between the most and least deprived male patients aged either 50, 65, or 80 years old at diagnosis.

```
foreach age in 50 65 80 {
 2.   rcsgen, scalar(`age´) knots($knots) rmatrix(Rage) gen(c) center(45)

 3.   global cancer_mostdep_`age´_male sex_cancer 0 sex_cvd 0 sex_other 0 ///
>         mostdep_cancer 1  mostdep_cvd 0  mostdep_other 0   ///
>         rcsage1_cancer `=c1´ rcsage2_cancer `=c2´ rcsage3_cancer `=c3´ ///
>         rcsage1_other 0 rcsage2_other 0 rcsage3_other 0 ///
>         rcsage1_cvd 0 rcsage2_cvd 0 rcsage3_cvd 0 _cancer 1 _cvd 0 _other 0

 4.   global cancer_leastdep_`age´_male sex_cancer 0 sex_cvd 0 sex_other 0 ///
>         mostdep_cancer 0 mostdep_cvd 0 mostdep_other 0 ///
>         rcsage1_cancer `=c1´ rcsage2_cancer `=c2´ rcsage3_cancer `=c3´ ///
>         rcsage1_other 0 rcsage2_other 0 rcsage3_other 0 ///
>         rcsage1_cvd 0 rcsage2_cvd 0 rcsage3_cvd 0 _cancer 1 _cvd 0 _other 0
 5. }
```

As we do not average over each observation, we must tell `standsurv` to only take the first observation in the stacked data to calculate non-marginalised predictions. This is done using `if _n == 1`. The `failure` option is used to obtain the cumulative incidence functions that is specified in each `at` option. To calculate the difference between `at1` and `at2`, we use `contrast(difference)`.

```
range tempt 0 10 101

foreach age in 50 65 80 {
 2.   foreach cause in cancer other cvd {
 3.      standsurv if _n==1, at1(${`cause´_leastdep_`age´_male}) ///
>         at2(${`cause´_mostdep_`age´_male}) ///
>         atvars(Fage`age´_`cause´_male_least CIF_`age´_`cause´_male_most) ///
>         contrastvar(CIF_`age´_`cause´_male_diff) ///
>         contrast(difference) failure timevar(tempt) ci
```

```
        4.  }
        5. }
```

Since we are making predictions at particular covariate patterns for each of the causes separately, specifying `rmft` gives us estimates of the expected life-years lost due to a particular cause of death. To calculate RMLT, we need to take the sum of all of the `at` options, where the expected life-years lost due to cancer, CVD and other causes is specified in each. We do this by creating our own contrast in a user-defined mata function which can be called in the option `userfunction`. An example of this is also given below.

```
      foreach age in 50 65 80 {
       2.   foreach cause in cancer cvd other {
       3.      standsurv if _n==1, at1(${`cause´_leastdep_`age´_male}) ///
      >            at2(${`cause´_mostdep_`age´_male}) ///
      >            atvars(LYL_`age´_`cause´_leastdep LYL_`age´_`cause´_mostdep) ///
      >            contrast(difference) contrastvar(LYL_`cause´`age´_diff) ///
      >            rmft timevar(tempt) ci
       4.   }
       5. }

      mata mata clear
      mata
      function RMFT(at) {
     2.   return((at[1]:+at[2]:+at[3]))
     3. }

      end
```

In order to obtain marginalised estimates, in each `at` option, only the covariate pattern for the group of interest need to be given. For the covariate distribution that we want to average over, as we have created interactions between the covariates and the causes of death, these must be mapped to each covariate e.g. `sex_cancer = sex`. The others are excluded from the `at` option for the other causes of death. In this case, because we want to average over covariates that we wish to standardise by, we need to identify the row for each patient in the stacked data that corresponds

to the failure time of that individual. This is done by creating the indicator variable `first` and using it as an `if` condition in `standsurv`. As before, we give an example for specifying macros for use in the `at` options for deaths due to cancer.

```
. global cancer_mostdep_stand sex_cvd 0 sex_other 0 sex_cancer = sex ///
>   mostdep_cancer 1  mostdep_cvd 0 mostdep_other 0 ///
>   rcsage1_cancer = rcsage1 rcsage2_cancer = rcsage2 rcsage3_cancer = rcsage3 ///
>   rcsage1_other 0 rcsage2_other 0 rcsage3_other 0 ///
>   rcsage1_cvd 0 rcsage2_cvd 0 rcsage3_cvd 0 ///
>   _cancer 1 _cvd 0 _other 1

. global cancer_leastdep_stand sex_cvd 0 sex_other 0 sex_cancer = sex ///
>   mostdep_cancer 0  mostdep_cvd 0 mostdep_other 0 ///
>   rcsage1_cancer = rcsage1 rcsage2_cancer = rcsage2 rcsage3_cancer = rcsage3 ///
>   rcsage1_other 0 rcsage2_other 0 rcsage3_other 0 ///
>   rcsage1_cvd 0 rcsage2_cvd 0 rcsage3_cvd 0 ///
>   _cancer 1 _cvd 0 _other 1

. bysort failcode id (_t): gen first = _n==1
```

The cause-specific CIF differences are thus calculated as follows,

```
. foreach cause in cancer other cvd {
  2.   standsurv if first, at1(${`cause´_leastdep_stand}) ///
>         at2(${`cause´_mostdep_stand}) ///
>         atvars(Fstand_`cause´_least Fstand_`cause´_most) ///
>         contrast(difference) contrastvars(Fdiff_`cause´) ///
>         failure timevar(tempt) ci
  3. }
```

As highlighted above, we can write user-functions to define our own contrasts. Below is an example for when interest is in calculating the difference in RMLT between the most and least deprived patients.

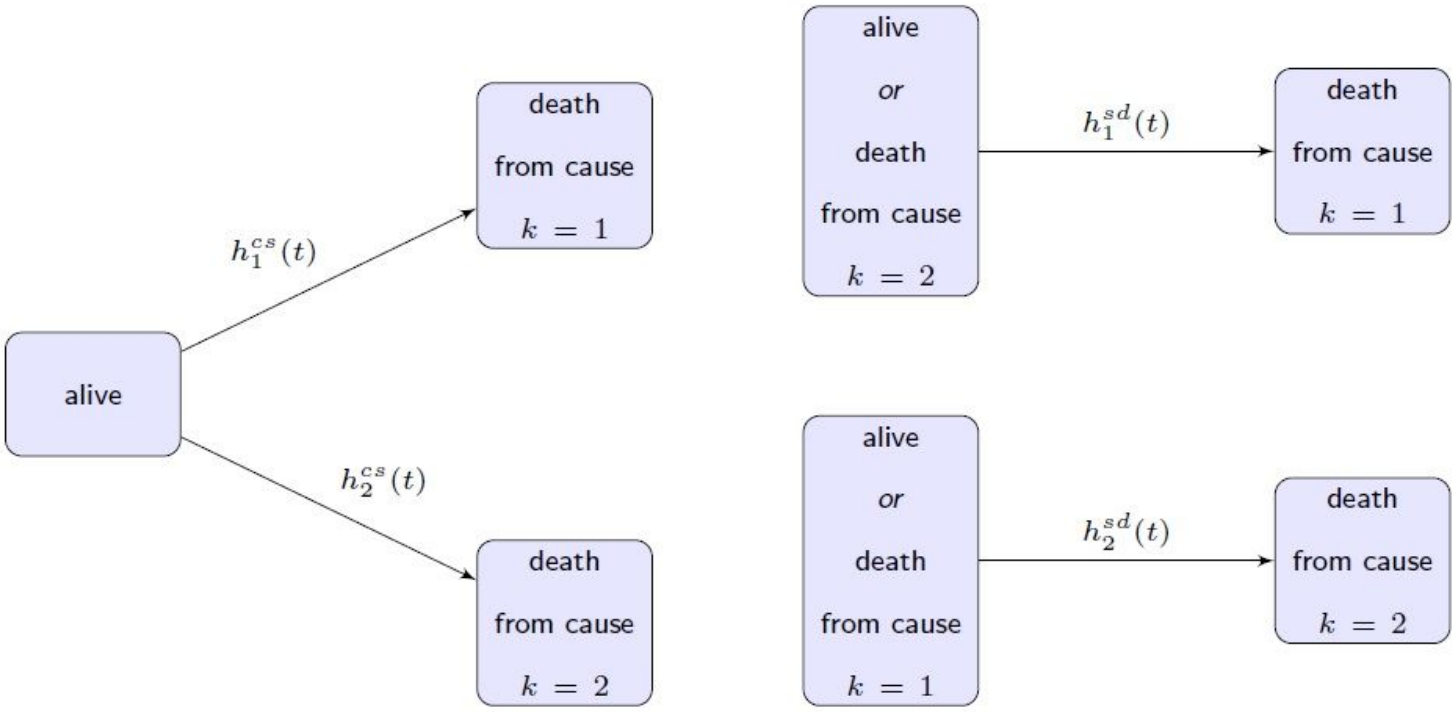
```
. mata mata clear
. mata
: function RMFTdiff(at) {
2.   return((at[1]:+at[2]:+at[3]) :- (at[4]:+at[5]:+at[6]))
3. }

: end

. standsurv if first, at1(${cancer_mostdep_stand}) ///
```

```
992         >   at2(${cvd_mostdep_stand}) ///
993         >   at3(${other_mostdep_stand}) ///
994         >   at4(${cancer_leastdep_stand}) ///
995         >   at5(${cvd_leastdep_stand}) ///
996         >   at6(${other_leastdep_stand}) ///
997         >   atvars(LYLcancer_stand_mostdep LYLcvd_stand_mostdep ///
998         >   LYLother_stand_mostdep LYLcancer_stand_leastdep ///
999         >   LYLcvd_stand_leastdep LYLother_stand_leastdep) ///
1000        >   userfunction(RMFTdiff) userfunctionvar(RMFT_diff) ///
1001        >   failure timevar(tempt) ci
```

# Figures

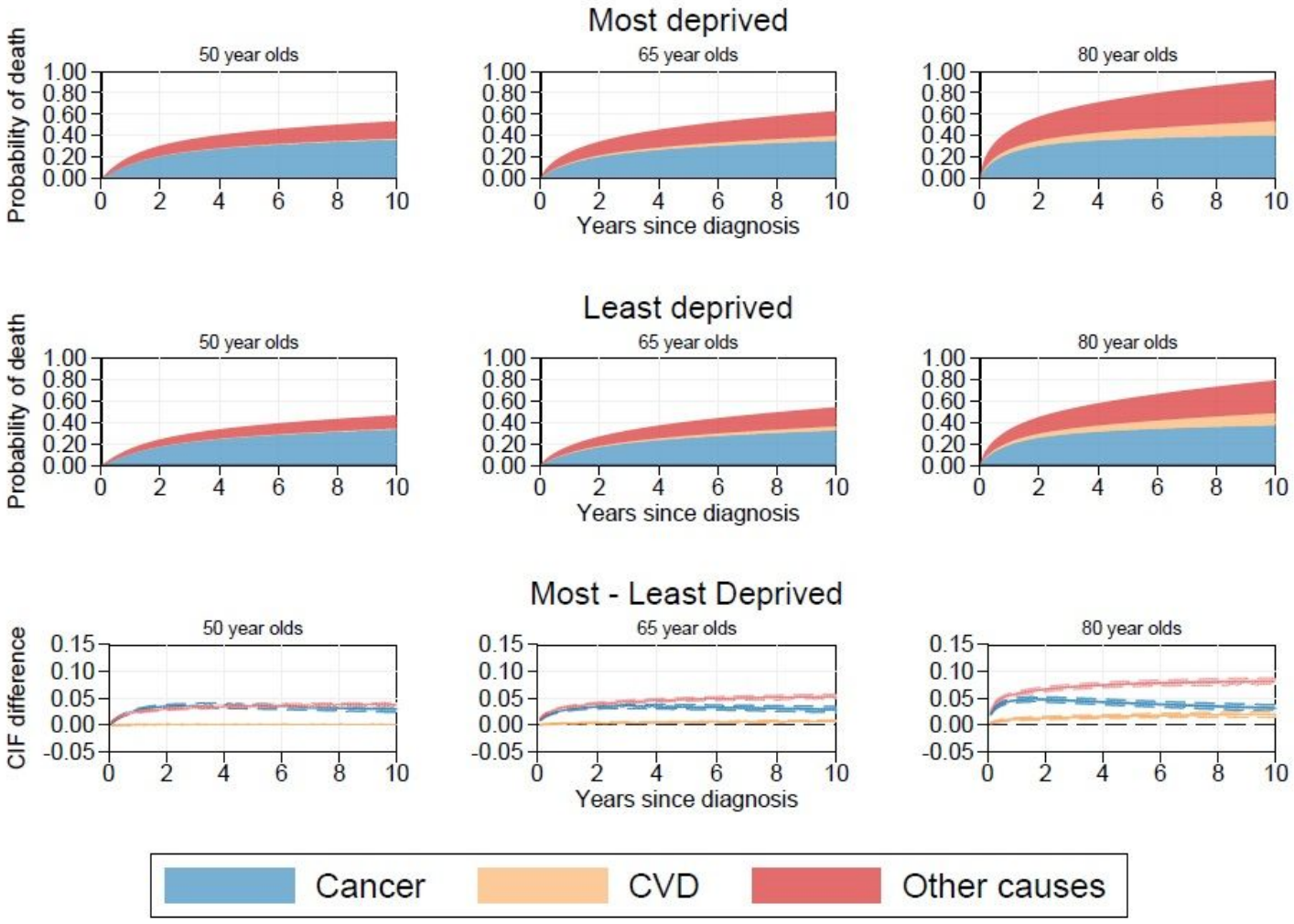



(a) Transitions described by the cause-specific hazard rate. The typical competing risks scenario.

(b) Transitions described by the subdistribution hazard rate.

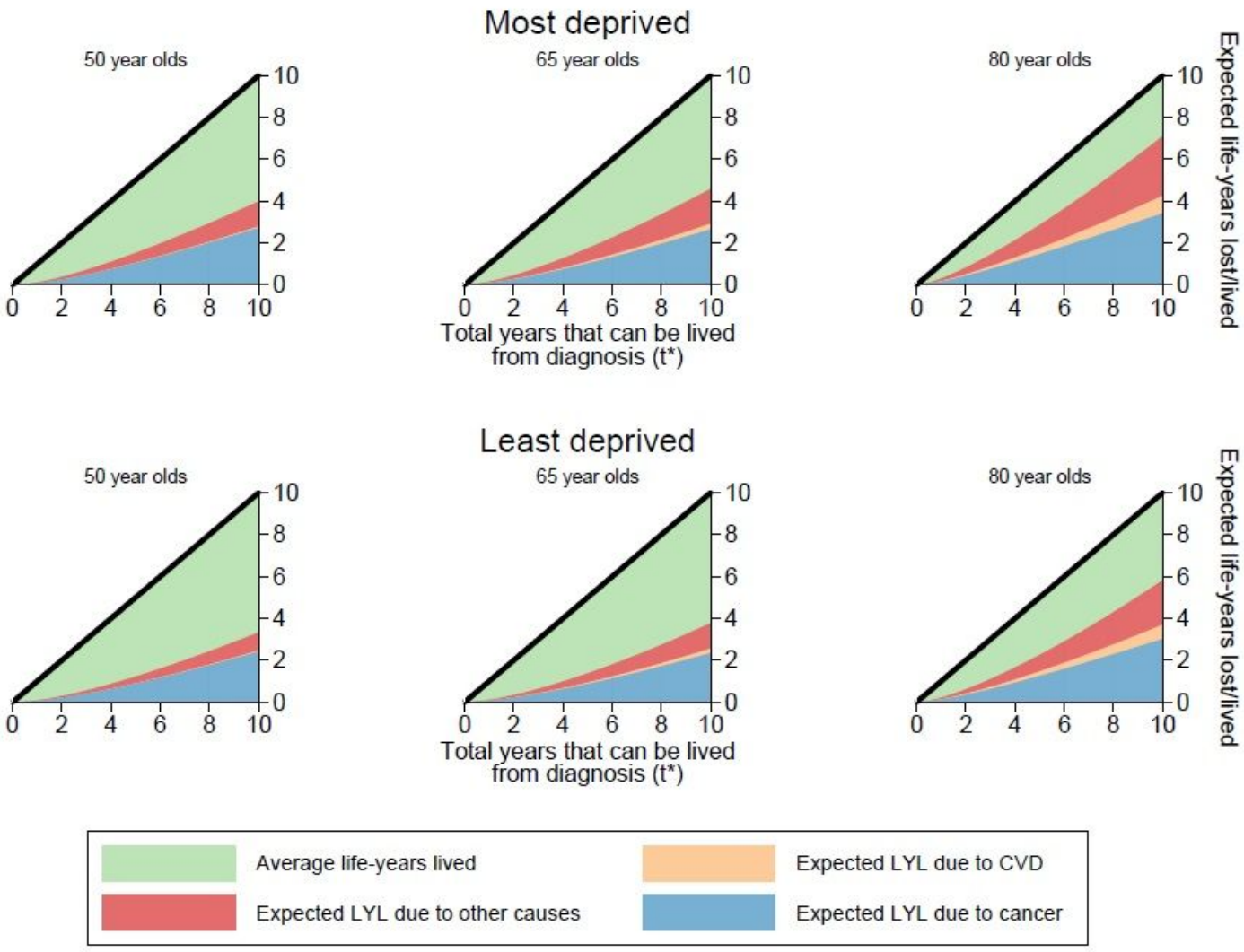
**Figure 1**

Schematics of transitions from an initial state to one of K = 2 causes of death.




## Figure 2

Stacked cause-specic CIFs by deprivation group and CIF dif ferences for male patients at specic ages.




## Figure 3

Stacked plots of expected life-years lost partitioned by each cause of death for male patients.

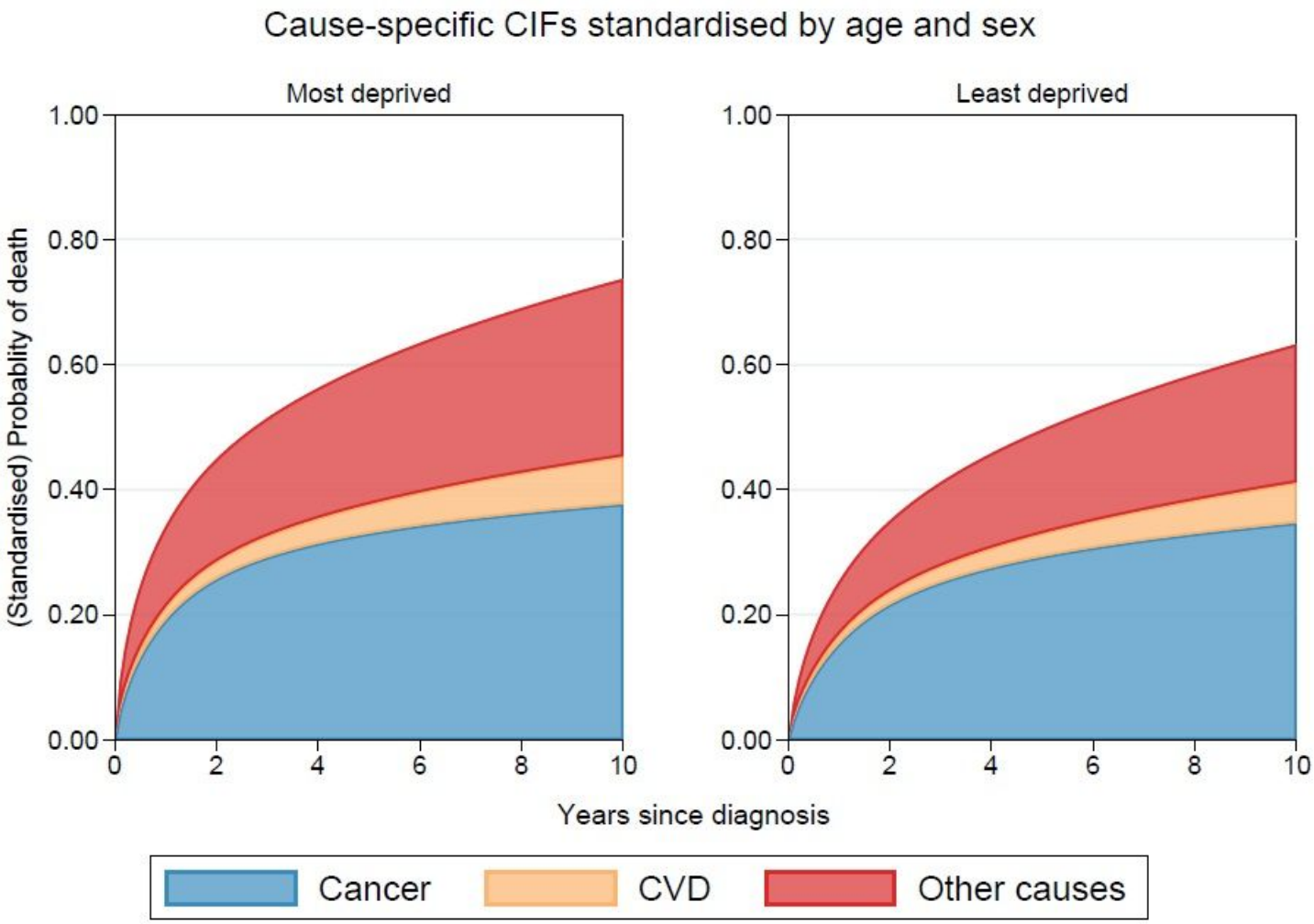
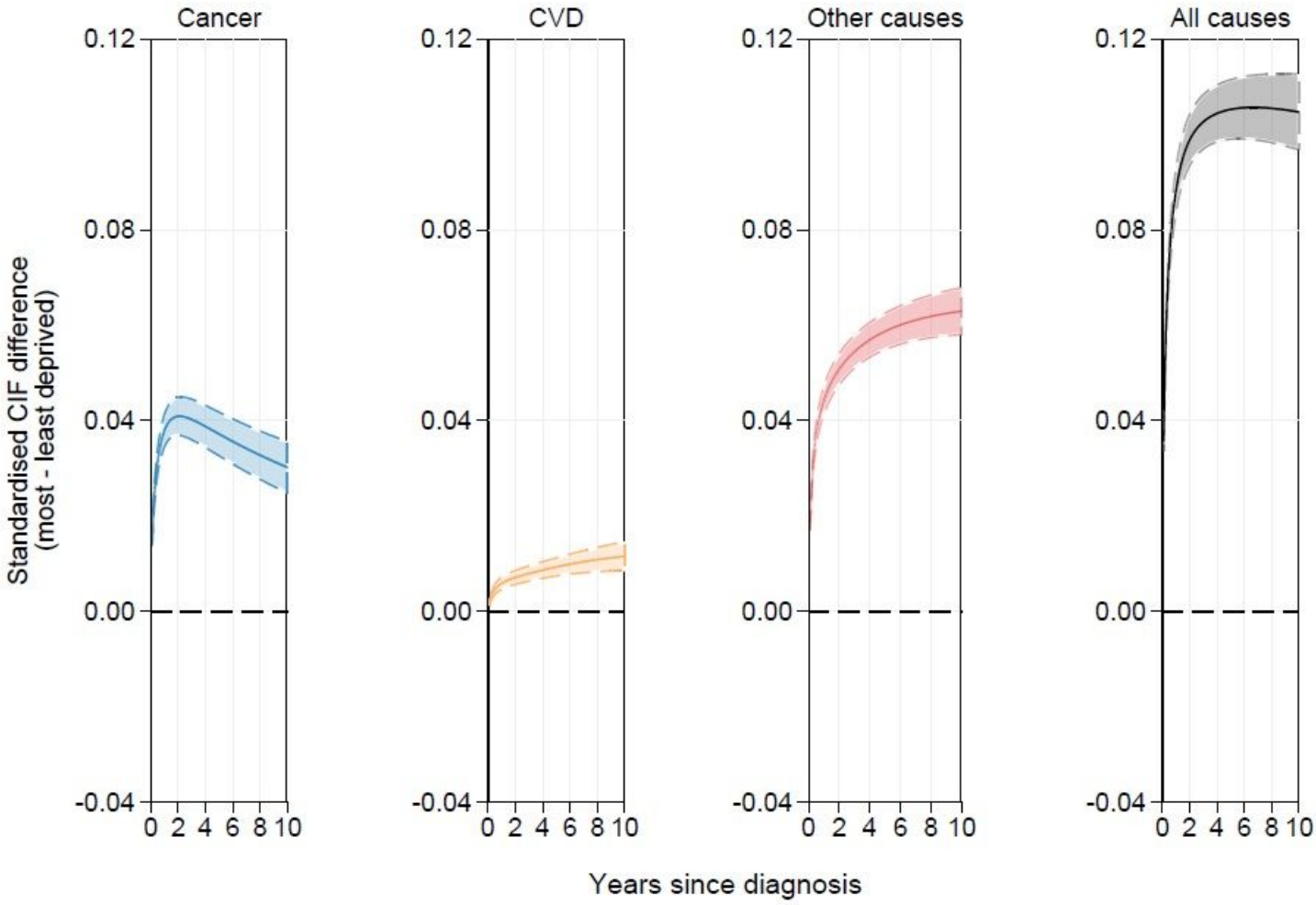



**Figure 4**

Estimated cause-specic CIFs standardised by age and sex for each deprivation group.




## Figure 5

Estimated CIF dif frences for each cause of death standard- ised by age and sex with 95% CIs.

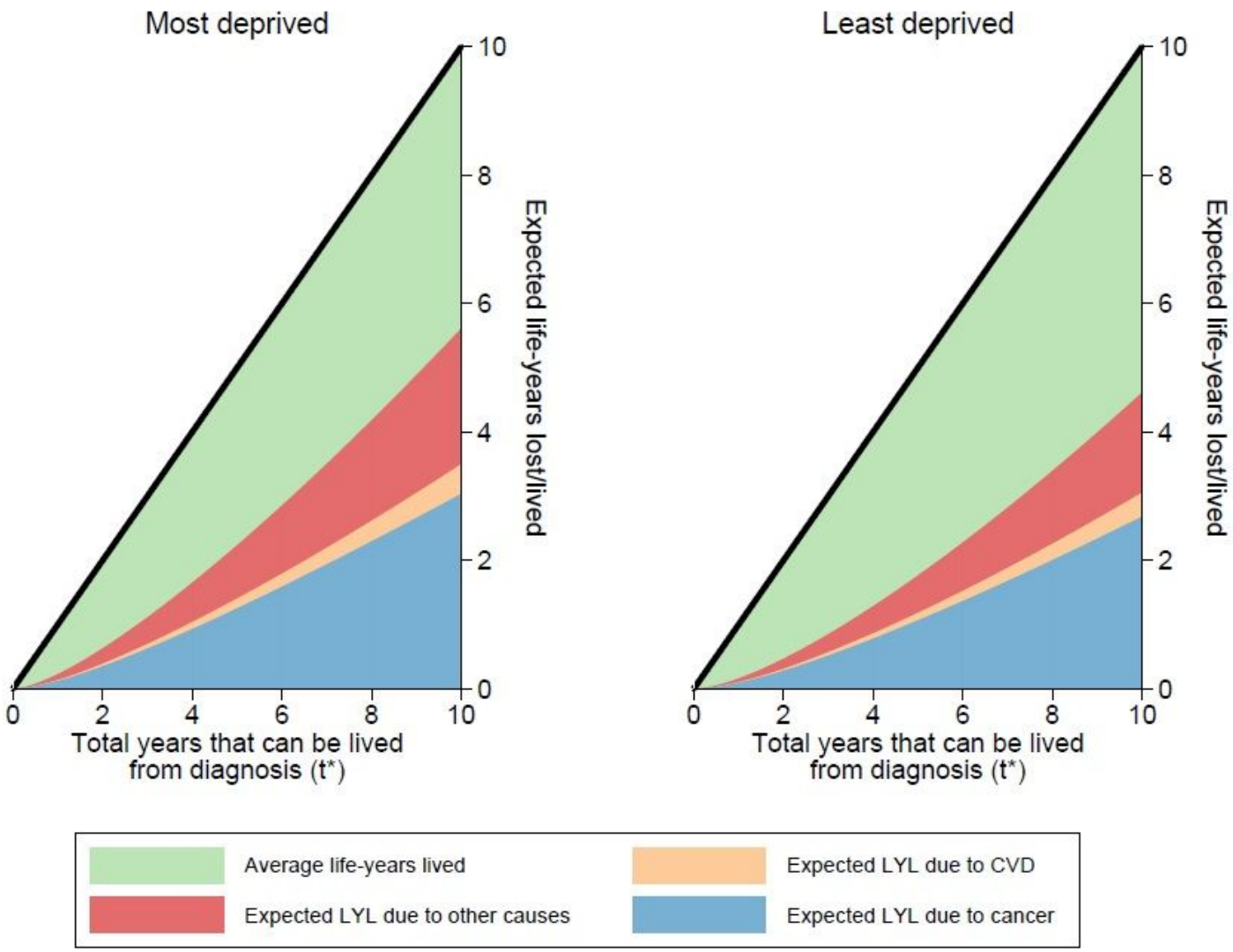



**Figure 6**

Stacked plots of expected life-years lost for each cause of death standardised by age and sex.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- stset.log.tex
- stcrprep.log.tex
- bmcart.cls
- bmcmathphys.bst
- rmftstand2.log.tex
- stata.sty
- model2.log.tex

- rmftage.log.tex
- covstand2.log.tex
- model.log.tex
- covage.log.tex
- cifstand2.log.tex
- cifage.log.tex
- codeappendix.tex
- refs02Feb19.bib