Direct likelihood inference on the cause-specific cumulative incidence function: a flexible parametric regression modelling approach

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

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

E-mail for correspondence: si113@le.ac.uk

In a competing risks analysis, interest lies in the cause-specific cumulative incidence function (CIF) which can be calculated by either (1) transforming on the cause-specific hazard (CSH) or (2) through its direct relationship with the subdistribution hazard (SDH). We expand on current competing risks methodology from within the flexible parametric survival modelling framework (FPM) and focus on approach (2). This models all cause-specific CIFs simultaneously and is more useful when we look to questions on prognosis. We also extend cure models using a similar approach described by Andersson *et. al.* for flexible parametric relative survival models. Using SEER public use colorectal data, we compare and contrast our approach with standard methods such as the Fine & Gray model, and show that many useful out-of-sample predictions can be made after modelling the cause-specific CIFs using a FPM approach. Alternative link functions may also be incorporated such as the logit link. Models can also be easily extended for time-dependent effects.

Introduction

To understand more about patient prognosis and disease impact, the probability of death due to a particular cause in the presence of other causes is needed and involves the consideration of competing causes of death. This probability is known as the cause-specific cumulative incidence function (CIF). From a statistical modelling perspective, this is usually obtained by either (1) estimating all the cause-specific hazard (CSH) functions, or (2) transforming using a direct relationship with the subdistribution hazard (SDH) function for the cause of interest. The choice of model on which to make our statistical inference depends on the research question to be answered. Wolbers *et. al.*,

At present, the most commonly implemented method for modelling covariate effects on the cause-specific CIF is the Fine & Gray model for the SDH.

Over a reasonably long enough follow up time, the cause-specific CIF for most cancers reach a plateau, referred to as "statistical" cure. At this point, patients no longer die from the cancer of interest and instead die from the other competing events, in which case, modelling the cure proportion amongst cancer patients may be of interest.

In this paper, we introduce a FPM approach for direct likelihood inference on the cause-specific CIF. The model is then extended to estimate the cure proportion and estimate the probability of patients bound to die from cancer amongst those that are still alive.

Methodology

Let T be the time to event any of K competing causes $k = 1, \cdots, K$ and D denote the type of event, where $D = 1, \cdots, K$. Here, we consider the events to be death from different causes and so the cause-specific CIF, $F_k(t)$, is the probability of dying from a particular cause, $D = k$, by time t whilst also being at risk of dying from other causes

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

The all-cause CIF, $F(t)$, which is the probability of dying from any of the K causes by time t, is the sum of all K cause-specific CIFs and can also be expressed as the complement of the overall survival function, $S(t)$,

$$F(t) = P(T \leq t) = \sum_{j=1}^{K} F_j(t) = 1 - S(t) \tag{2}$$

Cause-specific and subdistribution hazard functions

The cause-specific CIF, $F_k(t)$, can be expressed as a function of the SDH for cause k or expressed as a function of the CSH functions for all k causes. The CSH function, $h_k^{cs}(t)$ gives 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.

$$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k | T > t)}{\Delta t} \tag{3}$$

The cause-specific CIF can be written as a function of the CSHs for all K causes such that,

$$F_k(t) = \int_0^t \left(\exp\left[-\int_0^t \sum_{j=1}^{K} h_j^{cs}(u)du \right] \right) h_k^{cs}(u)du \tag{4}$$

Note here that the leading term within the integral gives the overall survival function, $S(t)$.
Gray

$$h_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k | T > t \cup (T \leq t \cap D \neq k)}{\Delta t} = \frac{\frac{d}{dt}\left[F_k(t)\right]}{1 - F_k(t)} = -\frac{d\left[\ln\left(1 - F_k(t)\right)\right]}{dt} \tag{5}$$

and is interpreted as the instantaneous 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.

$$F_k(t) = 1 - \exp\left[-H_k^{sd}(u)\right] \qquad \text{and} \qquad H_k^{sd}(t) = \int_0^t h_k^{sd}(u)du \tag{6}$$

An important distinction between the CSH and SDH for cause k is found within the risk-set. The risk-set in the CSHs is described in the conventional epidemiological sense, i.e. those who have died from any of the k causes of death, are no longer considered to be at risk. In contrast, the risk-set for the SDH for cause k considers patients who have died from any of the $K-1$ competing causes, excluding cause k, to still be at risk from dying of the cause of interest, k. A more detailed description and comparison of the risk-sets for the CSH and SDH for cause k can be found in Lau *et. al.*

A useful relationship between the SDH and CSH was highlighted by Beyersmann and Schumacher

$$h_k^{cs}(t) = h_k^{sd}(t) \left[1 + \frac{\left[\sum_{j=1}^{K} F_j(z)\right] - F_k(t)}{1 - \sum_{j=1}^{K} F_j(t)} \right] \tag{7}$$

Thus using the SDH functions for all K causes, we can also obtain the CSH functions for all K causes.

0.1 Regression modelling

A common approach for modelling the CSH function is by assuming proportional hazards (PH) using the Cox model. So with covariates, $\mathbf{x}$, we have that,

$$h_k^{cs}(t|\mathbf{x}) = h_{0,k}^{cs}(t) \exp\left[\mathbf{x}\boldsymbol{\beta}_k^{cs}\right] \tag{8}$$

where $\boldsymbol{\beta}_k^{cs}(t)$ are log cause-specific hazard ratios (HR), and $h_{0,k}^{cs}$ is the baseline CSH function. To re-iterate, in order to estimate one cause-specific CIF, it is necessary to estimate the CSHs for all k causes (see Equation 4).

Alternatively, the most common model for the SDH for cause k is the Fine & Gray model.

$$h_k^{sd}(t|\mathbf{x}) = h_{0,k}^{sd}(t) \exp\left[\mathbf{x}\boldsymbol{\beta}_k^{sd}\right] \tag{9}$$

where $\boldsymbol{\beta}_k^{sd}$ are log-SDH ratios (SHR) for cause k.

A key difference between the two regression models in Equation 8 and Equation 9 is in the interpretation of the parameters $\exp\left(\boldsymbol{\beta}_k^{cs}\right)$ (HRs) and $\exp\left(\boldsymbol{\beta}_k^{sd}\right)$ (SHRs). The HRs give us the association on the effect of a covariate on the cause-specific mortality rate and SHRs give the association on the effect of a covariate on risk (refer to Wolbers *et. al.*

Likelihood estimation

We first describe parametric inference on K competing causes of death under the CSH approach, which models using the standard survival likelihood function with an observable failure or censoring time, t_i, with independent and non-informative right censoring, for each individual $i = 1, \cdots, N$,

$$L = \prod_{i=1}^{N} \left[\left[\prod_{j=1}^{K} \left[S(t_i|\mathbf{x}_i) h_j^{cs}(t_i|\mathbf{x}_i)\right]^{\delta_{ij}} \right] \left[S(t_i|\mathbf{x}_i)\right]^{1 - \sum_{j=1}^{K} \delta_{ij}} \right] \tag{10}$$

where the censoring indicator, δ_{ik}, tell us whether an individual died from any cause k ($\delta_{ik} = 1$), or not ($\delta_{ik} = 0$) and $S(t_i|\mathbf{x}_k)$ is the overall survival function.

Alternatively, Jeong and Fine

$$L = \prod_{i=1}^{N} \left[\left[\prod_{j=1}^{K} \left[h_j^{sd}(t_i|\mathbf{x}_i)(1 - F_j(t_i|\mathbf{x}_i)) \right]^{\delta_{ij}} \right] \left[1 - \sum_{j=1}^{K} F_j(t_i|\mathbf{x}_i) \right]^{1-\sum_{j=1}^{K} \delta_{ij}} \right] \tag{11}$$

Note here, however, that, the cause-specific CIF, $F_k(t)$, in Equation 11 is not a proper cumulative distribution function and is instead referred to as a subdistribution function since $\lim_{t\to\infty} F_k(t) < 1$.

Flexible parametric regression model for the cause-specific CIF

Like the Cox model, the Fine & Gray model estimates covariate effects but does not specifically model the underlying baseline rates. We propose a parametric survival model which directly estimates both the covariate effects on the cause-specific CIF and the underlying baseline using the likelihood in Equation 11 simultaneously for all K causes. Standard parametric models such as the exponential, Weibull or Gompertz distributions, are often unable to capture more complex underlying baseline hazard functions containing one or more turning points.

$$\ln(H_k^{sd}(t)) = g(F_k(t|\mathbf{x}_{ik})) = s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\boldsymbol{\beta}_k = \gamma_{0k} + \gamma_{1k}z_{1k} + \cdots + \gamma_{(M-1)k}z_{(M-1)k} + \mathbf{x}_k\boldsymbol{\beta}_k \tag{12}$$

Where $z_{1k}, \cdots, z_{(M-1)k}$ are the basis functions of the RCS and are defined as follows:

$$z_{1k} = \ln(t) \tag{13}$$

$$z_{jk} = (\ln(t) - m_{jk})_+^3 - \phi_{jk}(\ln(t) - m_{1k})_+^3 - (1 - \phi_{jk})(\ln(t) - n_{Mk})_+^3, \qquad j = 2, \cdots, M-1$$

where,

$$\phi_{jk} = \frac{n_{Mk} - n_{jk}}{n_{Mk} - n_{1k}} \tag{14}$$

and

$$(u)_+ = \begin{cases} u, & \text{if } u < 0 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

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. The choice of the position and number of knots is subjective, which is used as an argument for a drawback of the flexible parametric modelling framework. However, others have explored this through a variety of sensitivity analyses of the knots and it has been shown to have very little influence on obtained predictions (please refer to Hinchliffe and Lambert

Link functions

We showed in Equation 12 that we can derive a log-cumulative SDH model with covariates and through the general link function, $g(\cdot)$, for the cause-specific CIF, $F_k(t)$, are able to apply similar transformations described in Royston and Parmar

The majority of regression models are specified through the complementary log-log link function which we will mainly focus on in this paper,

$$g(F_k(t|\mathbf{x}_{ik})) = \ln\left[-\ln(1 - F_k(t|\mathbf{x}_k))\right] \tag{16}$$

and we can calculate the SDH function for each cause k and the cause-specific CIF, which are defined as follows,

$$h_k^{sd}(t|\mathbf{x}_k) = \frac{d\left[s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k)\right]}{dt} \exp\left(s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\boldsymbol{\beta}_k\right) \tag{17}$$

$$F_k(t|\mathbf{x}_k) = 1 - \exp\left(-\exp\left[s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\boldsymbol{\beta}_k\right]\right) \tag{18}$$

where the $\boldsymbol{\beta}_k$'s are log-SHRs.

Alternatively, Gerds *et. al.*

$$g(F_k(t|\mathbf{x}_k)) = \text{logit}\left(F_k(t|\mathbf{x}_k)\right) \tag{19}$$

and the cause-specific CIF is,

$$F_k(t|\mathbf{x}_k) = \frac{\exp\left[s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\boldsymbol{\beta}_k\right]}{1 + \exp\left[s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\beta_k\right]} \tag{20}$$

The logit link model described above, describes the probability of dying from the competing cause k in relation to the probability of not experiencing the competing event k which includes those that are still alive and those that have already died from one of the other competing events. Gerds *et. al.*

Time-dependent effects to model non-proportionality

In Section 0.1, using the link function in Equation 16, we defined a proportional log-cumulative SDH FPM with RCS for the underlying baseline log-cumulative baseline SDH simultaneously for all K causes. A natural advantage of these models is that we can easily extend to incorporate time-dependent effects to model non-proportionality. This is achieved by fitting interactions between the associated covariates and the spline functions. Using this interaction, we can introduce a new set of knots, $\mathbf{m}_{ek}$, which represent the eth time-dependent effect for cause k with associated parameters $\boldsymbol{\alpha}_{ek}$. If there are $e = 1, \cdots, E$ time-dependent effects, we can extend the cause-specific log-cumulative SDH in Equation 25 to,

$$\ln(H_k^{sd}(t)) = \eta_k(t) = s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_{0k}) + \mathbf{x}_k\boldsymbol{\beta}_k + \sum_{l=1}^{E} s_k(\ln(t); \boldsymbol{\alpha}_{lk}, \mathbf{m}_{lk})x_{lk} \tag{21}$$

In this approach, the spline function for different time-dependent effects can be different and requires fewer knots to the baseline spline function.[?] This is an extension on the original approach proposed by Royston and Parmar.

Estimating the cure proportion

Andersson *et. al.*

$$F_c(t|\mathbf{x}_c) = 1 - (1 - \pi_c)^{\exp\left[\gamma_{2c}z_{2c} + \cdots + \gamma_{(M-1)c}z_{(M-1)c} + \sum_{i=1}^{E} s_c(\ln(t); \boldsymbol{\alpha}_{ic}, \mathbf{m}_{ic})\mathbf{x}_{ic}\right]} \tag{22}$$

where,

$$1 - \pi_c = 1 - \exp\left(-\exp(\gamma_{0c} + \mathbf{x}_c\boldsymbol{\beta}_c)\right) \tag{23}$$

Therefore, the constant parameters, γ_{0c} and $\mathbf{x}_c$ are used to model the cure proportion for cause $k = c$. Here, we also implement a constraint on the linear spline, γ_{1c}, such that it is equal 0.

A useful prediction from these models is the estimate of the proportion of patients that will eventually die, or are bound-to-die, from cancer, or other causes, of those that are still alive. It should be noted, however, that this is a measure at the population level and individual patients are not specified to a particular group. Where a plateau is observed for a particular cause, e.g. cancer, the cancer-specific CIF will no longer increase beyond a given point in time and allows estimation of the proportion of patients bound-to-die of cancer amongst those that are still alive.

$$P_{alive,can}(t) = \pi_c - F_1(t) \tag{24}$$

$$P_{alive,oth}(t) = 1 - F_2(t) - \cdots - F_K(t) - \pi_c \tag{25}$$

where π_c is the proportion of those bound-to-die from cancer on which cure is assumed. These are a useful summary measure of patient prognosis and further complements the direct FPMs on the cause-specific CIFs when interest primarily lies in answering more prognostic-related research questions.

Simulation

Modelling the SDH for a particular cause is usually performed using the Fine & Gray approach and is widely considered as the standard for modelling covariate effects directly on the cause-specific CIF. To contrast this approach against the log-cumulative subdistribution hazards (-CSDH) FPM, we carried out a simple simulation to demonstrate that, like the Fine & Gray model, unbiased estimates are also obtained with good coverage. Furthermore, we will present relative gain in precision (RPG) of modelling under the log-CSDH FPM over the Fine & Gray approach. We also aim to explore a common area of concern around the use of log-CSDH FPM which is in the choice of the number of knots, or degrees of freedom, for the restricted cubic splines. In our simulation results, we hope to echo what has already been shown in previous simulation studies regarding the use of restricted cubic splines in flexible parametric survival models.[?]

Design

A simulation study was designed with one scenario where true SDH functions were generated for two causes which approached an asymptote of 0. These SDH functions were chosen to demonstrate that restricted cubic splines were robust enough to handle a scenario when there is potential for the optimiser to search in the incorrect direction leading to negative SDHs. The complexity in the shape of SDH functions for both causes were formulated under the mixture Weibull distribution with the assumption of proportionality.

The design of the simulation study is outlined below,

1. Survival times were generated from CSH functions for both causes which were transformed from SDH functions generated from mixture Weibull distributions using the relationship in equation 7.[?] The shape, γ, scale, λ and mixture, p, parameters were chosen such that the SDH functions for both causes tended to an asymptote of 0 (see Figure **??**). SDH functions were generated under the assumption of proportionality between the covariate groups for each cause using simulated competing risks data based on CSH functions as derived by Beyersmann *et. al.*

Results

Table **??** summarises the obtained log-SHRs for cause 1 and standard errors from 1000 replicated datasets with 200, 500 and 5000 observations. The simulation under the above parameters generated a mean of 22% right-censored individuals for 200 and 5000 observations and 23% for 500 observations and a mean of 28% failures from cause 1 for 200, 500 and 5000 observations. The bias, i.e. difference between the model log-SHR and true log-SHR of -0.5, coverage and rMSE is given. Overall, for the models that converge, it is clear that under both the Fine & Gray and FPM approach, we get negligible bias, indicating that all models, irrespective of the number of degrees of freedom used for the baseline RCS, are unbiased. We also demonstrate good coverage in all of the models. Finally, a marginally lower rMSE is observed in all of the log-CSDH FPMs in comparison to the Fine & Gray approach. This demonstrates that, overall, estimates are obtained with a lower bias and more precision under the FPM approach over the standard method.

Similarly, also in Table **??**, we have the bias, coverage and rMSE for the cause-specific CIFs at 1, 3 and 5 years since diagnosis. Again, we show that there is negligible bias in the estimates from all the models, good coverage is consistently shown over time and we also have similar rMSE across all the models. Overall, the simulation shows that, regardless of the number of degrees of freedom used for the baseline RCS, the parameters are stable across all the models and any differences between them are negligible.

However, convergence issues arise in the smaller simulated datasets for 200 and 500 observations. Non-convergence especially arise when more complicated models are fitted i.e. more degrees of freedoms are used. This suggests potential over-fitting of the models to the data since, for example, using 3 to 4 degrees of freedom in the simulation with 500 observations leads to no problems in model convergence.

Illustrative Example

In this Section, we provide an example to illustrate the different predictions available after fitting a FPM to directly model all cause-specific CIFs. We further demonstrate that we can more accurately capture the shape of the data when fitting FPMs with time-dependent effects to relax the assumption of proportionality. The prediction of other useful predictions to aid interpretation in these more complex models are also demonstrated.

Description of data

We demonstrate the methodology outlined in this paper through the use of SEER public use colorectal data.

Proportional subdistribution hazards models

Separate Fine & Gray models were fitted for each of the 3 causes with stage at diagnosis as the only covariate. To illustrate the estimation process, we initially restricted analysis to patients aged above 75 years old where competing risks are more likely to make an impact. Parameter estimates are compared against those predicted under the FPM approach which were fitted for the log-cumulative baseline SDH for all 3 causes simultaneously using 5 df for the baseline RCS. Table **??** shows the fitted estimates from a Fine & Gray model and a log-CSDH FPM. The apparent disagreement between the estimated subdistribution hazard ratios (SHRs) and their 95% CIs can be partially explained by the unreasonable assumption of proportionality of the effect of stage at diagnosis for all 3 causes being made on the competing causes in the FPM approach. More complex models were fitted in order to demonstrate this issue by fitting 3 separate log-CSDH FPMs by including time-dependent effects for the other competing events. For this data, because there is non-proportionality, it is accounted for by including time-dependent effects for all causes when modelling using the FPM approach, which is more sensible (see Section 0.1). These "adjusted" estimates are also compared in Table **??** which is labelled Log-CSDH FPM2 and good agreement between all SHRs and their 95% CI is now observed. The estimated cause-specific CIFs from both models are illustrated in Figure **??**. Here, it is clear that the Fine & Gray Model and Log-CSDH FPM2 yield similar estimates and we observe very good agreement between the two curves.

Non-proportional subdistribution hazards models

Generally, the effect of stage on mortality is stronger shortly after diagnosis compared to later on in time, indicating that proportional SDH may not be a reasonable assumption. To relax this assumption, time-dependent effects are included to allow the effect of stage at diagnosis to vary over time for all K causes of death using RCS with 3 df. To assess the accuracy in estimation, predictions from the model are compared to empirical estimates of the SDH for cause k using the Aalen-Johansen estimator for the cause-specific CIF

Transforming to the cause-specific hazards

From these log-CSDH regression models, we are also able to estimate the CSH functions since we model the SDH functions for all K causes using Equation 7. We return to analysing the full dataset and in Figure **??**, the CSHs derived from a standard flexible parametric CSH regression model, as described by Hinchliffe and Lambert

Other useful predictions

The advantage of fitting FPMs and modelling all K causes simultaneously is that it is easy to obtain other predictions which aids interpretation. For example, as shown in Figure **??**, we can present absolute differences in the cause-specific CIF for 65 year olds between covariate groups. 95% CIs can be calculated for these measures using the delta method

Cure models

In order to fit cure models, it must be reasonable to assume cure on the observed dataset. To assess the appropriateness of the cure assumption for cancer, the Aalen-Johansen empirical estimates were compared against the cancer-specific CIFs estimated from a log-CSDH cure model. Analysis was restricted to patients with regional stage cancer at diagnosis and the youngest age group, i.e. 55 to 64 year olds, where cure is found to be a reasonable assumption. Cure was modelled for patients who died from colorectal cancer and 5 df were used for the baseline RCS in the log-CSDH FPM. From Figure **??**, it can be seen that, after approximately 13 years, the empirical curve plateaus at just above 30% and in comparison, the cancer-specific CIF predicted from the model slightly underestimates the cure proportion. Over follow-up time a good agreement is observed between the Aalen-Johansen and model estimates and overall, cure appears to be reasonable. Useful predictions are also estimable from the cure model such as the proportion of patients who are bound to die from cancer, or other causes, of those that are alive. The plot to the right in Figure **??** represents the stacked probabilities for each cause-specific CIF. The cancer-specific CIF plateaus at about 12 years after diagnosis and the cure proportion is estimated at 30%. The dashed-line partitions those who are still alive into two groups. For example, at 3 years after diagnosis, 20% have died and 15% are alive and bound to die from cancer and 65% are alive and not bound to die from cancer. However, at approximately 12 years since diagnosis, as the point of cure is approached, it is expected that, about 40% of patients have died and the remaining 60% are almost all bound to die from causes other than colorectal cancer.

Discussion

In this paper, extending on the ideas of Jeong and Fine

In general, modelling covariate effects on the cause-specific CIF in large population based studies requires relaxing the proportionality assumption. Including time-dependent effects in the FPM approach to relax the proportionality assumption is much quicker and less computationally intensive as there is no need to incorporate time-dependent weights on an expanded dataset or fit separate models for each of the cause-specific CIFs

In contrast to the Fine & Gray model, researchers are able to model all cause-specific CIFs simultaneously, which is better for a deeper understanding of the effects of covariates on all cause-specific CIFs and allows us to answer more complicated research questions on patient prognosis. Although these models may be more complex, there are a number of useful estimable measures including absolute differences in the cause-specific CIFs and relative contributions to the total mortality. Accompanied with these predictions, using the delta method, we can also obtain 95% CIs. Although it is also theoretically possible to obtain CIs for predictions from the Fine & Gray model, in practise, this is computationally intensive and is usually done using bootstrapping methods which is not optimal for large datasets. Hence, modelling using the approach in this paper is more accessible and easier to implement for researchers, especially when analysing larger datasets in the hundreds of thousands.

Even though our approach estimates parameter effects on the cumulative incidence, because all cause-specific CIFs are modelled together, we show that the CSH functions can be estimated. However, since the multiplier in this equation

is time-dependent, although the assumption of proportionality may be reasonable on the SDH scale, this may not also be true on the CSHs and vice versa

Another useful property of simultaneously modelling all cause-specific CIFs in a direct likelihood FPM approach, is that the methods can be easily extended to model the cure proportion. The method for estimating cure described by Andersson *et. al.*

Limitations

A well-known problem of direct regression models for the cause-specific CIF is that the sum of all probabilities may exceed 1 for certain covariate patterns. This is particularly problematic in the oldest age groups where patients are at a higher risk of dying from competing events leading to very high overall probability of death. This is also the case in our approach and it is sometimes avoided if models are not misspecified, for example, by adjusting for all appropriate covariates with any potential interactions and by including time-dependent effects. In some situations models may fail to converge when specified correctly, but this will depend on the use of better initial values for the optimiser so that it is not searching in the wrong direction. As an informal assessment of misspecification of the models, we can compare the CSHs derived from our approach to standard CSH regression modelling methods by allowing for appropriate model complexity on both scales. However, in many datasets, the all-cause CIF will not get close to one, since, in many studies, follow-up is usually restricted. Shi *et. al.*

If interest is only in the covariate effects on one cause, it is not imperative to model all cause-specific CIFs as this may unnecessarily complicate the analysis. In these cases, a single Fine & Gray model may suffice or model the cause-specific CIF using time-dependent weights

A potential criticism of the FPM approach is the need to specify the positioning and number of knots. However, this has been shown to have little influence on the cause-specific CIF through sensitivity analyses and other similar studies have also been carried out on the sensitivity of knots

In smaller simulated datasets, where $N = 200, 500$, some models struggled to converge under the FPM approach. In these cases, since the likelihood is evaluated at the last observed time for either cause, we found that the reason for non-convergence was mainly attributed down to insufficient follow-up time for a cause which led to inappropriate extrapolation. Other possible reasons for convergence issues in these smaller datasets, as mentioned previously, may be due to the lack of events for a given cause towards the last observed follow-up time and over-fitting models. Therefore, when fitting FPMs to smaller data, such as clinical trial data, it is recommended that fewer degrees of freedom are used for the restricted cubic splines. However, this paper concentrates on the implementation of methods for population-based data which usually contain observations well above 5000. Hence, as demonstrated in the simulation, fitting models using the FPM approach in this scenario show excellent performance regardless of the choice in the number of degrees of freedom.

Conclusions

The choice of analytic approach ultimately depends on the research question to be answered and the scale on which we wish to make our inferences. Our proposed method is most useful when we wish to make inferences on absolute risks and understand covariate effects on all of the cause-specific CIFs simultaneously. As discussed above, there are further advantages of implementation from within a FPM approach. A generalisation of the Weibull distribution with RCS is used to model and more flexibly capture the baseline log-cumulative SDH function. As opposed to standard semi-parametric approaches, since the cumulative SDH function is estimated in FPMs, it is easy to obtain other predictions that facilitate risk communication, some of which have already been discussed. Alternatively, to make inferences on aetiology, the alternative CSH approach for FPMs is available, making it possible to fit equivalent models on both scales. In fact, literature suggests that reporting on both CSH and SDH regression models is advantageous for understanding the overall impact of cancer on risk. CSH functions are also easy to derive from the flexible parametric SDH regression models in this paper since all K causes are modelled simultaneously. Finally, to ensure that the methods are accessible for researchers, a user-friendly command, `stpm2cr`, is available in Stata

Online Appendix

```
// Fit PSDH FPM
do dataimport.do
// stset data with all-causes
stset survmm, failure(cause == 1, 2, 3) scale(12) id(id) exit(time 180)
stpm2cr [colrec_cancer: stage2, scale(hazard) df(5)] ///
```

```
[other_causes: stage2, scale(hazard) df(5)] ///
[heart_disease: stage2, scale(hazard) df(5)] ///
, events(cause) cause(1 2 3) cens(0) eform mlmethod(lf2) old

range _t1 `=1/12´ 15 1000

// Make out-of-sample cause-specific cif predictions for each covariate pattern
predict direct_cif_loc, cif at(stage1 1 stage2 0) timevar(_t1)
predict direct_cif_reg, cif at(stage1 0 stage2 1) timevar(_t1)

// Fit PSDH FPM2
do dataimport.do

stset survmm, failure(cause == 1, 2, 3) scale(12) id(id) exit(time 180)

// Fit 3 seperate models for each cause assuming proportionality, and time dependent effects
//      for the competing causes.
*Cancer
stpm2cr [colrec_cancer: stage2, scale(hazard) df(5)] ///
[other_causes: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
[heart_disease: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
, events(cause) cause(1 2 3) cens(0) eform mlmethod(lf2) old

. range _t2 `=1/12´ 15 1000

predict direct_cif_adj_loc, cif at(stage1 1 stage2 0) cause(1) timevar(_t2)
predict direct_cif_adj_reg, cif at(stage1 0 stage2 1) cause(1) timevar(_t2)

*Other Causes
stpm2cr [other_causes: stage2, scale(hazard) df(5)] ///
[colrec_cancer: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
[heart_disease: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
, events(cause) cause(2 1 3) cens(0) eform mlmethod(lf2) old

predict direct_cif_adj_loc, cif at(stage1 1 stage2 0) cause(2) timevar(_t2)
predict direct_cif_adj_reg, cif at(stage1 0 stage2 1) cause(2) timevar(_t2)

*Heart Disease
stpm2cr [heart_disease: stage2, scale(hazard) df(5)] ///
[colrec_cancer: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
[other_causes: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
, events(cause) cause(3 1 2) cens(0) eform mlmethod(lf2) old

predict direct_cif_adj_loc, cif at(stage1 1 stage2 0) cause(3) timevar(_t2)
predict direct_cif_adj_reg, cif at(stage1 0 stage2 1) cause(3)  timevar(_t2)

// Fit Non-PSDH FPM
do dataimport.do

stset survmm, failure(cause == 1, 2, 3) scale(12) id(id) exit(time 180)

stpm2cr [colrec_cancer: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
[other_causes: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
[heart_disease: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
, events(cause) cause(1 2 3) cens(0) eform mlmethod(lf2) old

range _t3 `=1/12´ 15 1000

predict directtvc_cif_loc, cif at(stage1 1 stage2 0) timevar(_t3)
predict directtvc_cif_reg, cif at(stage1 0 stage2 1) timevar(_t3)

/* Direct Model for each of the 3 causes adjusted for stage and age */
do dataimport.do

stset survmm, failure(cause == 1, 2, 3) scale(12) id(id) exit(time 180)

// Generate RCS with 3 df to allow for non-linear effects on age at diagnosis
rcsgen age, gen(rcsage) df(3) orthog

global knots `r(knots)´

matrix Rage = r(R)

// Fit the log-cumulative subdistribution hazard FPM
stpm2cr [colrec_cancer: stage2 rcsage1-rcsage3, scale(hazard) df(5) tvc(rcsage1-rcsage3 stage2) dftvc(3)] ///
[other_causes: stage2 rcsage1-rcsage3, scale(hazard) df(5) tvc(rcsage1-rcsage3 stage2) dftvc(3)] ///
[heart_disease: stage2 rcsage1-rcsage3, scale(hazard) df(5) tvc(rcsage1-rcsage3 stage2) dftvc(3)] ///
, events(cause) cause(1 2 3) cens(0) eform mlmethod(lf2) old

range _t2 `=1/12´ 15 1000

// Obtain out-of-sample predictions at individual ages for patients
forvalues age = 60(10)80 {
// Generate and store splines for a particular age for use in predict
rcsgen , scalar(`age´) knots($knots) rmatrix(Rage) gen(c)

// Predict cause-specific cumulative incidence functions
predict cif_loc`age´, cif at(stage1 1 stage2 0 rcsage1 `=c1´ rcsage2 `=c2´ rcsage3 `=c3´) timevar(_t2)
predict cif_reg`age´, cif at(stage1 0 stage2 1 rcsage1 `=c1´ rcsage2 `=c2´ rcsage3 `=c3´) timevar(_t2)

// Predict subdistribution hazards
predict sdh_loc`age´, sub at(stage1 1 stage2 0 rcsage1 `=c1´ rcsage2 `=c2´ rcsage3 `=c3´) timevar(_t2)
predict sdh_reg`age´, sub at(stage1 0 stage2 1 rcsage1 `=c1´ rcsage2 `=c2´ rcsage3 `=c3´) timevar(_t2)
```

```
// Predict cause-specific hazards
predict csh_loc`age´, csh at(stage1 1 stage2 0 rcsage1 `=c1´ rcsage2 `=c2´ rcsage3 `=c3´) timevar(_t2)
predict csh_reg`age´, csh at(stage1 0 stage2 1 rcsage1 `=c1´ rcsage2 `=c2´ rcsage3 `=c3´) timevar(_t2)

// Predict subdistribution hazard ratios
predict shr_reg`age´, shrn(stage1 0 stage2 1 rcsage1 `=c1´ rcsage2 `=c2´ rcsage3 `=c3´) ///
shrd(stage1 1 stage2 0 rcsage1 `=c1´ rcsage2 `=c2´ rcsage3 `=c3´) timevar(_t2)

// Predict cause-specific hazard ratios
predict chr_reg`age´, chrn(stage1 0 stage2 1 rcsage1 `=c1´ rcsage2 `=c2´ rcsage3 `=c3´) ///
chrd(stage1 1 stage2 0 rcsage1 `=c1´ rcsage2 `=c2´ rcsage3 `=c3´) timevar(_t2)
}
```

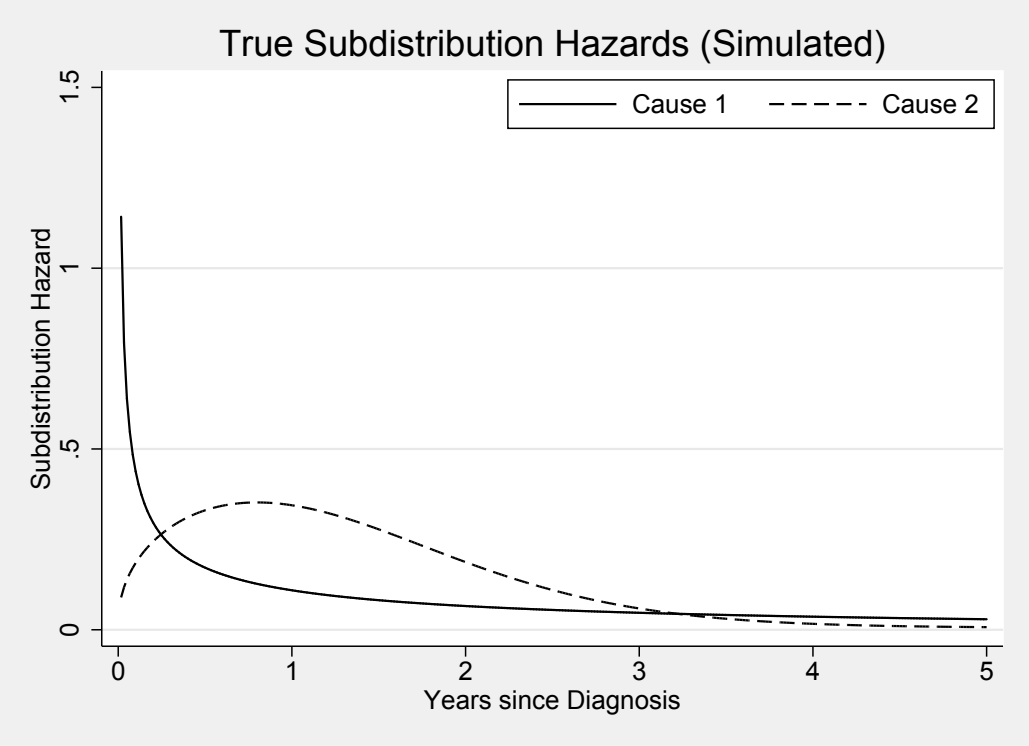

Figure 1: Subdistribution hazards (SDH) simulated from a mixture Weibull distribution with paramaters $\lambda_{1,1} = 0.6$, $\gamma_{1,1} = 0.5$, $\lambda_{1,2} = 0.01$, $\gamma_{1,2} = 0.35$ and $p_1 = 0.5$ for the SDH for cause 1 and $\lambda_{2,1} = 0.01$, $\gamma_{2,1} = 0.8$, $\lambda_{2,2} = 0.7$, $\gamma_{2,2} = 1.45$ and $p_2 = 0.5$ for cause 2

Table 1: Simulation results for the log-subdistribution hazard ratio (SHR) and cause-specific cumulative incidence functio
proportional subdistribution hazards models with two competing causes and one binary covariate X.

			Log-SHR = -0.5			CIF for Cause 1					
						Year 1			Year 3		
N	Code	Converged (%)	Bias	Coverage	rMSE	Bias	Coverage	rMSE	Bias	Coverage	rMSE
200	fg	100.0	-0.0295	0.9640	0.2691	-0.0004	0.9420	0.0383	-0.0004	0.9530	0.0464
	3df	95.7	-0.0248	0.9613	0.2645	0.0018	0.9592	0.0369	-0.0020	0.9540	0.0447
	4df	96.0	-0.0266	0.9625	0.2633	0.0003	0.9542	0.0375	-0.0000	0.9573	0.0451
	5df	92.2	-0.0287	0.9642	0.2633	0.0016	0.9469	0.0382	-0.0006	0.9501	0.0454
	6df	93.0	-0.0255	0.9613	0.2639	0.0002	0.9473	0.0378	-0.0007	0.9473	0.0456
	9df	79.2	-0.0262	0.9684	0.2592	0.0012	0.9470	0.0378	0.0012	0.9558	0.0458
500	fg	100.0	-0.0111	0.9600	0.1697	-0.0009	0.9620	0.0248	-0.0008	0.9540	0.0297
	3df	100.0	-0.0101	0.9600	0.1668	0.0017	0.9560	0.0241	-0.0024	0.9600	0.0282
	4df	100.0	-0.0115	0.9600	0.1674	0.0001	0.9500	0.0246	-0.0002	0.9580	0.0284
	5df	99.0	-0.0130	0.9596	0.1675	0.0008	0.9556	0.0247	-0.0008	0.9616	0.0285
	6df	97.4	-0.0124	0.9589	0.1683	-0.0003	0.9610	0.0244	-0.0003	0.9610	0.0289
	9df	97.8	-0.0127	0.9611	0.1680	-0.0003	0.9591	0.0245	-0.0004	0.9611	0.0290
5000	fg	100.0	0.0012	0.9570	0.0550	-0.0009	0.9560	0.0077	-0.0009	0.9450	0.0097
	3df	100.0	0.0027	0.9560	0.0531	0.0009	0.9750	0.0073	-0.0026	0.9510	0.0094
	4df	100.0	0.0014	0.9560	0.0532	-0.0007	0.9550	0.0074	-0.0007	0.9600	0.0092
	5df	100.0	0.0011	0.9560	0.0532	0.0001	0.9600	0.0075	-0.0015	0.9630	0.0093
	6df	100.0	0.0009	0.9560	0.0533	-0.0009	0.9600	0.0075	-0.0008	0.9600	0.0093
	9df	100.0	0.0008	0.9560	0.0533	-0.0008	0.9660	0.0076	-0.0007	0.9600	0.0094

Table 2: Subdistribution hazard ratios (SHRs) estimated from a Fine & Gray model, log-cumulative subdistribution hazards flexible parametric model (Log-CSDH FPM) and a Log-CSDH FPM adjusted for time-depedent effects on the competing events (Log-CSDH FPM2). SHRs compare regional stage patients to localised stage patients aged 75 to 84 years old assuming proportionality.

	Fine & Gray Model			Log-CSDH FPM			Log-CSDH FPM2		
	SHR	95% CI		SHR	95% CI		SHR	95% CI	
Colorectal:	3.503	[3.224	3.805]	3.429	[3.157	3.725]	3.504	[3.225	3.808]
Other Causes:	0.753	[0.703	0.806]	0.720	[0.673	0.771]	0.737	[0.689	0.789]
Heart Disease:	0.731	[0.661	0.807]	0.686	[0.622	0.757]	0.719	[0.651	0.794]

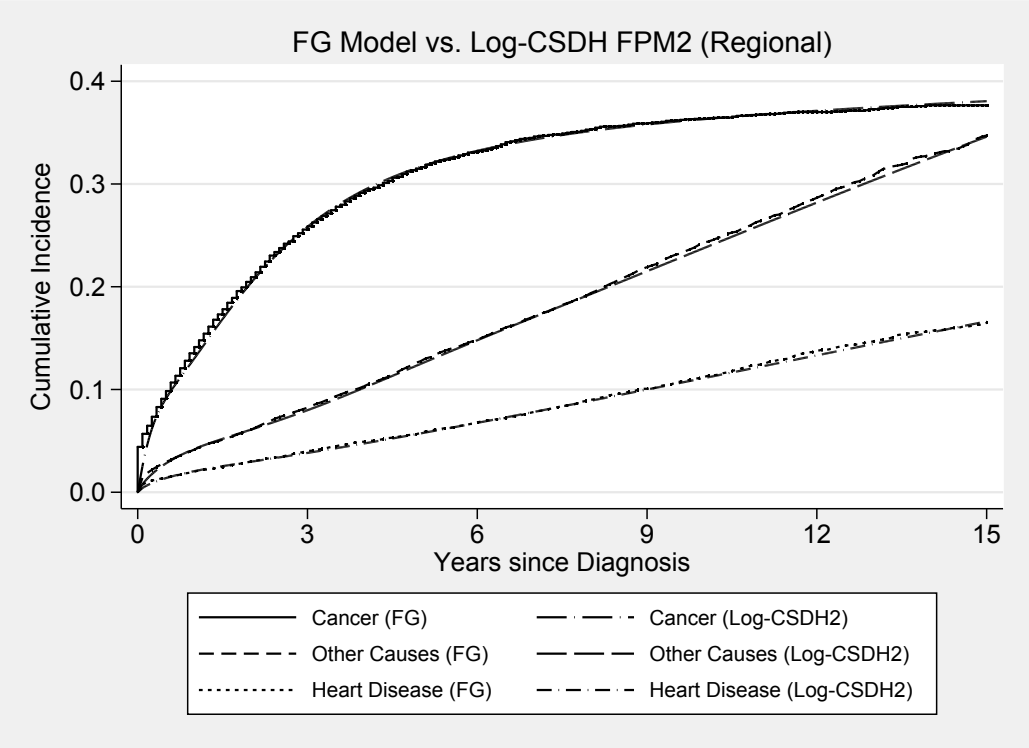

Figure 2: A comparison of predicted cause-specific cumulative incidence functions from a Fine & Gray (FG) model and a log-cumulative subdistribution hazards flexible parametric model adjusted for time-depedent effects on the competing events (Log-CSDH FPM2). Predictions are made for 75 to 84 year old male patients diagnosed with regional stage colorectal cancer.

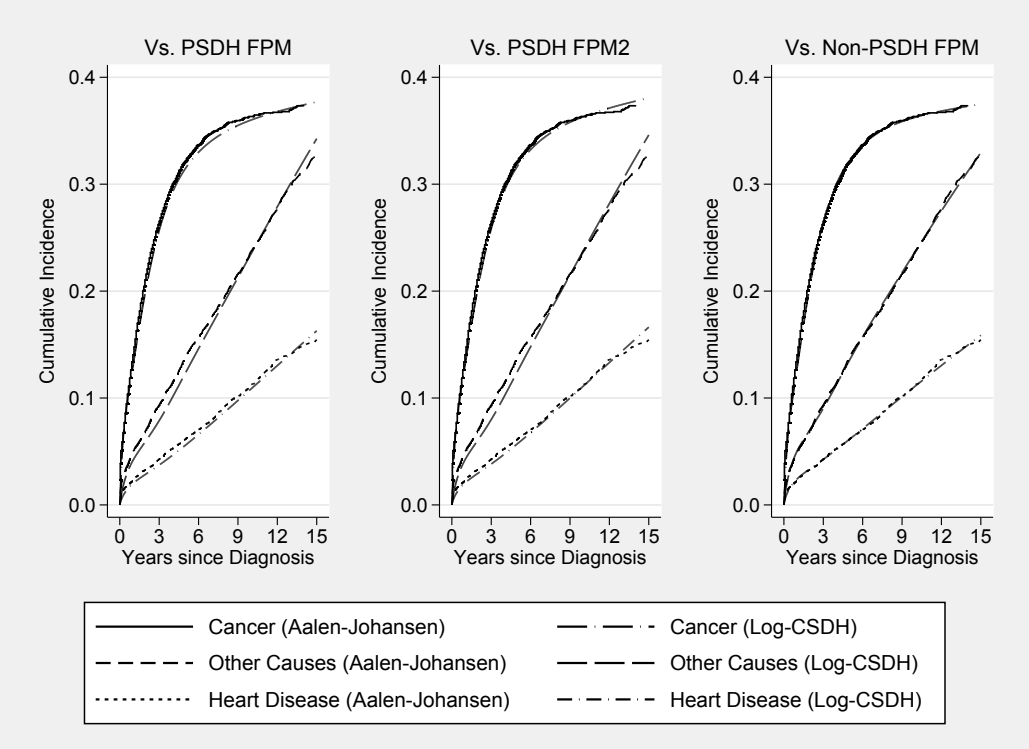

Figure 3: Predicted cause-specific cumulative incidence functions comparing empirical estimates (Aalen-Johansen) against a proportional log-cumulative subdistribution hazards flexible paramteric model adjusted for time-depedent effects on the competing events (PSDH FPM2) on the right plot and a non-proportional log-cumulative subdistribution hazards flexible parametric model (Non-PSDH FPM) on the left plot. Predictions are made for 75 to 84 year old male patients diagnosed with regional stage colorectal cancer.

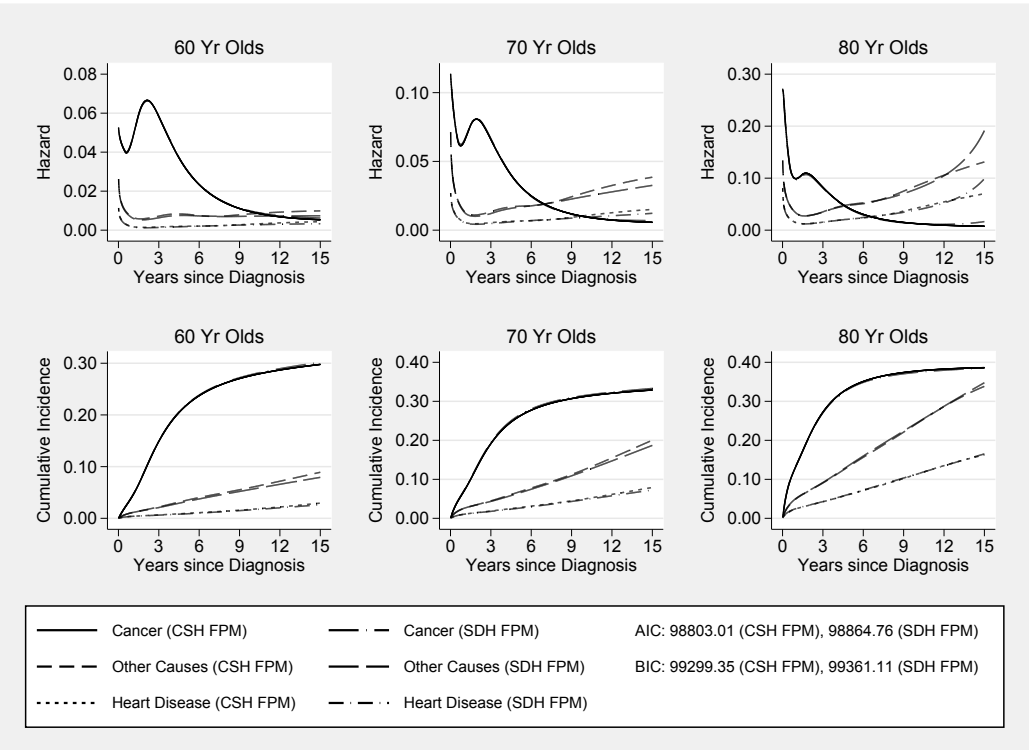

Figure 4: Predicted age-specific time-dependent cause-specific hazards and cumulative incidence functions after fitting a non-proportional log-cumulative subdistribution hazards flexible paramteric model (SDH FPM) comapred against a non-proportional cause-specific hazards flexible parametric model (CSH FPM) for 60, 70 and 80 year old male patients diagnosed with regional stage colorectal cancer.

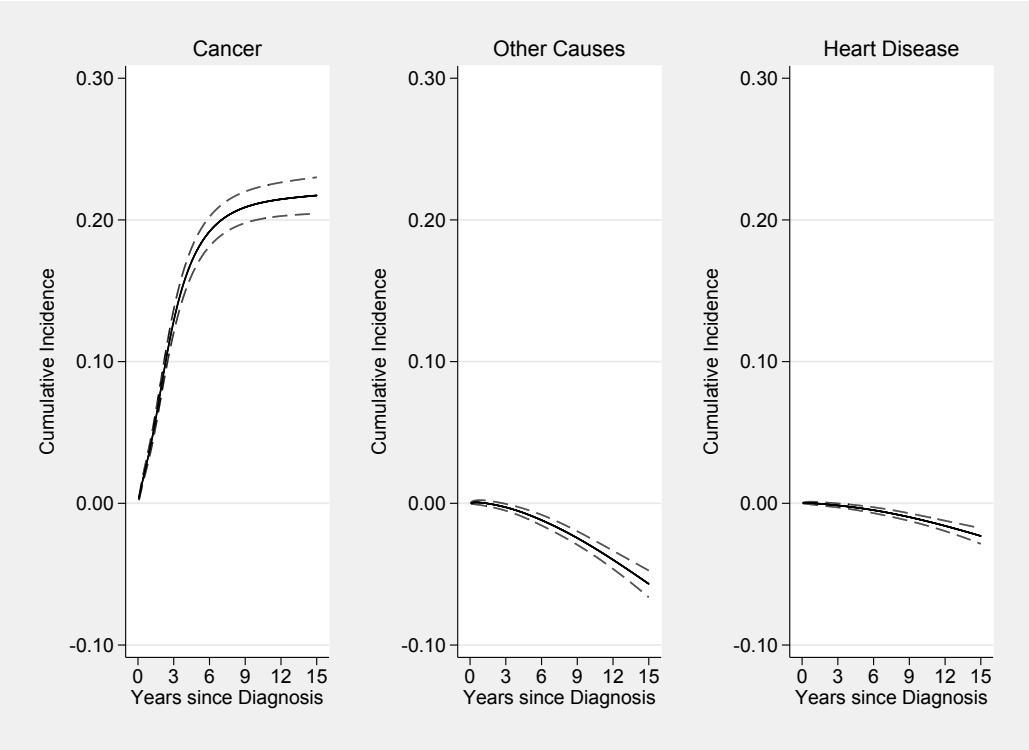

Figure 5: Absolute differences (regional stage minus localised stage), with 95% CIs (dashed line), between 65 year old patients with local and regional stage cancer at diagnosis.

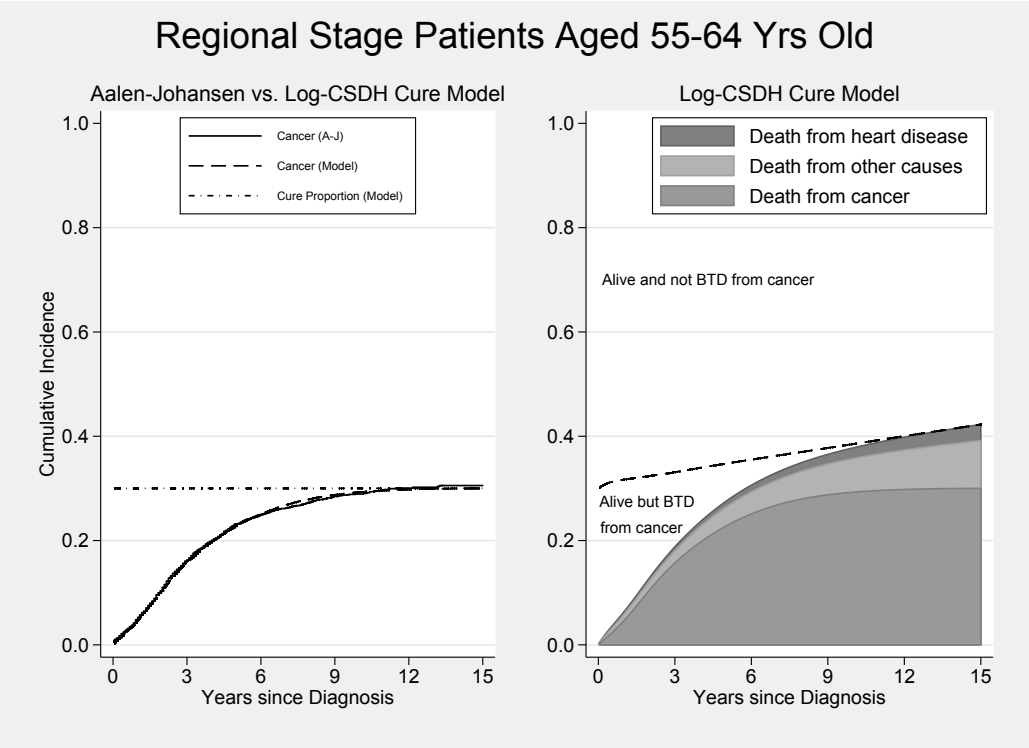

Figure 6: Predicted cancer-specific cumulative incidence functions (CIFs) for empirical Aalen-Johansen estimates compared against log-cumulative subdistribution hazards (Log-CSDH) estimates from a cure model (left). Stacked cause-specific CIFs and cure proportion (dashed-line) from a Log-CSDH cure model. The dashed-line partitions patients who are still alive into those who are bound to die (BTD) from cancer and not BTD from cancer (right). Predictions obtained for 55 to 64 year old male patients diagnosed with regional stage colorectal cancer.