

Using temporal recalibration to improve the calibration of risk prediction models in competing risk settings when there are trends in survival over time

Booth, Sarah; Mozumder, Sarwar I.; Archer, Lucinda; Ensor, Joie; Riley, Richard D.; Lambert, Paul C.; Rutherford, Mark J.

DOI:
[10.1002/sim.9898](https://doi.org/10.1002/sim.9898)

License:
Creative Commons: Attribution (CC BY)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Booth, S, Mozumder, SI, Archer, L, Ensor, J, Riley, RD, Lambert, PC & Rutherford, MJ 2023, 'Using temporal recalibration to improve the calibration of risk prediction models in competing risk settings when there are trends in survival over time', *Statistics in Medicine*. <https://doi.org/10.1002/sim.9898>

[Link to publication on Research at Birmingham portal](#)

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

RESEARCH ARTICLE

Using temporal recalibration to improve the calibration of risk prediction models in competing risk settings when there are trends in survival over time

Sarah Booth¹ | Sarwar I. Mozumder^{1,2} | Lucinda Archer³ | Joie Ensor³ |
Richard D. Riley³ | Paul C. Lambert^{1,4} | Mark J. Rutherford¹

¹Biostatistics Research Group,
Department of Population Health
Sciences, University of Leicester,
Leicester, UK

²Oncology Biometrics Statistical
Innovation, AstraZeneca, Cambridge, UK

³Institute of Applied Health Research,
College of Medical and Dental Sciences,
University of Birmingham, Birmingham,
UK

⁴Department of Medical Epidemiology
and Biostatistics, Karolinska Institutet,
Stockholm, Sweden

Correspondence

Sarah Booth, Biostatistics Research
Group, Department of Population Health
Sciences, University of Leicester,
Leicester, UK.

Email: sarah.booth@le.ac.uk

Funding information

Cancer Research UK, Grant/Award
Numbers: C14183/A29739,
C41379/A27583; Cancerfonden,
Grant/Award Number: 2018/744;
National Institute for Health and Care
Research, Grant/Award Number:
NIHR300100; UK Research and
Innovation; Vetenskapsrådet,
Grant/Award Number: 2017-01591

We have previously proposed temporal recalibration to account for trends in survival over time to improve the calibration of predictions from prognostic models for new patients. This involves first estimating the predictor effects using data from all individuals (full dataset) and then re-estimating the baseline using a subset of the most recent data whilst constraining the predictor effects to remain the same. In this article, we demonstrate how temporal recalibration can be applied in competing risk settings by recalibrating each cause-specific (or sub-distribution) hazard model separately. We illustrate this using an example of colon cancer survival with data from the Surveillance Epidemiology and End Results (SEER) program. Data from patients diagnosed in 1995–2004 were used to fit two models for deaths due to colon cancer and other causes respectively. We discuss considerations that need to be made in order to apply temporal recalibration such as the choice of data used in the recalibration step. We also demonstrate how to assess the calibration of these models in new data for patients diagnosed subsequently in 2005. Comparison was made to a standard analysis (when improvements over time are not taken into account) and a period analysis which is similar to temporal recalibration but differs in the data used to estimate the predictor effects. The 10-year calibration plots demonstrated that using the standard approach over-estimated the risk of death due to colon cancer and the total risk of death and that calibration was improved using temporal recalibration or period analysis.

KEYWORDS

calibration, competing risks, prognostic models, risk prediction, temporal recalibration

1 | INTRODUCTION

Developing prognostic models to produce long-term survival or risk predictions relies on the inclusion of patients diagnosed many years ago in order to have sufficient follow-up time to estimate the hazard rates, at for example, 10 years.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

However, if survival outcomes improve over time, and this is not accounted for during prognostic model development, then it can lead to out-dated risk predictions for new patients that overestimate the risk of death from different causes as well as the total risk of death. Previous external validation studies of risk prediction models have identified this type of calibration drift and stress the importance of monitoring model performance over time¹⁻⁵ and adjusting for miscalibration accordingly.⁶⁻⁸

In previous work, we proposed temporal recalibration as an approach to account for improvements in survival over time that can be applied at the model development stage and does not require any additional data.⁹ It is a two-step process where the model is first developed on the full dataset to estimate the predictor effects. The model is then recalibrated by re-estimating the (log cumulative) baseline hazard in a recent subsample of the data to update the baseline survival whilst constraining the predictor effects to remain the same as in the original model. This subsample is defined using a period window and the use of delayed entry allows improvements in baseline survival over time to be captured.

An alternative approach is to use period analysis¹⁰ to develop the prognostic model. Whilst in temporal recalibration the predictor effects are estimated using the full dataset, period analysis involves using just the subsample of data to estimate both the predictor effects and the baseline. This may lead to overfitting in derivation datasets with small numbers of participants or when the event of interest is rare.

Improvements in survival outcomes following a diagnosis of cancer have been widely reported over the past 20 years.¹¹⁻¹⁴ Using this knowledge during model development may lead to better calibrated risk predictions for new patients. We demonstrated this previously with a prognostic model for colon cancer patients, showing how temporal recalibration or period analysis can be used to account for trends in survival and lead to a prognostic model that produces better calibrated cause-specific survival predictions in comparison to using the standard method where these temporal trends are ignored.⁹

A further issue is that many prognostic model development situations involve competing risks, which should be accounted for, as if ignored this also leads to risk predictions that are too high.^{15,16} For example, when predicting the risk of death due to cancer, the competing risk of death due to other causes must be taken into account otherwise the predicted risk of death from cancer will be too high. In clinical practice, it may also be useful to understand not only a patient's risk of death from cancer but also their risk of death from other causes as this may affect treatment decisions,¹⁶ particularly for patients with comorbidities.

Established risk prediction models for cancer in England include PREDICT Breast,¹⁷ PREDICT Prostate,¹⁸ and Qcancer Colorectal Survival.¹⁹ These models produce long-term risk predictions that account for competing events by fitting a cause-specific hazard model for death due to the cancer of interest and a second model for death due to other causes. However, how to also consider the issue of changes in survival over time remains unaddressed.

In this article, we demonstrate how temporal recalibration can be extended into a competing risk setting, to allow for prognostic model development in situations with both trends in survival and competing risks. We focus on the cause-specific setting and show how to temporally recalibrate each of the cause-specific hazard models separately, which can then be combined to produce the risk predictions. An alternative analysis using the Fine and Gray approach is included in Appendix A.7. We illustrate the approach using an example of survival following a diagnosis of colon cancer, and provide further applications for lung and breast cancer in Appendices A.5 and A.6.

2 | METHODS

2.1 | Calculation of risk predictions

The key estimands of interest for prognostic models in competing risk settings are the cause-specific cumulative incidence functions (CIFs), $F_k(t|\mathbf{x}_i)$. These give the probability of failure (risk) due to cause k , by time point t , for an individual with covariate pattern \mathbf{x}_i , whilst accounting for the competing events. In the presence of competing risks, prognostic models can either be developed on the cause-specific hazard scale or alternatively the effects of predictors can be modelled directly on the CIFs using models on the subdistribution hazards scale for example, with Fine and Gray models.²⁰ Here we focus on the cause-specific setting but further discussion surrounding Fine and Gray models can be found in Section 2.2.

The cause-specific CIF for cause k , $F_k(t|\mathbf{x}_i)$, can be defined as follows:

$$F_k(t|\mathbf{x}_i) = \int_0^t S(u|\mathbf{x}_i) h_k(u|\mathbf{x}_i) du, \quad (1)$$

which depends on the cause-specific hazard function $h_k(t|\mathbf{x}_i)$ and the all-cause survival function $S(t|\mathbf{x}_i)$. The all-cause survival function is the product of the k cause-specific survival functions $S_k(t|\mathbf{x}_i)$ ^{15,21,22}:

$$S(t|\mathbf{x}_i) = \prod_{k=1}^K S_k(t|\mathbf{x}_i) = \exp\left(-\int_0^t \sum_{k=1}^K h_k(u|\mathbf{x}_i) du\right). \quad (2)$$

Each cause-specific survival function, $S_k(t|\mathbf{x}_i)$, can be defined with respect to the cause-specific hazard function $h_k(t|\mathbf{x}_i)$ as follows^{23,24}

$$S_k(t|\mathbf{x}_i) = \exp\left(-\int_0^t h_k(u|\mathbf{x}_i) du\right). \quad (3)$$

The integral for calculating the CIF will often not have a closed form solution and therefore may need to be obtained numerically.²³ However, the model parameters can be exported in order to produce risk predictions in a new dataset if the model were to be independently and externally validated (see Appendix A.1.4).

The total, or all-cause, risk of death at time t ($F_{all}(t|\mathbf{x}_i)$) can then be estimated as the sum of the k cause-specific CIFs or the complement of the all-cause survival function as shown by^{16,25}

$$F_{all}(t|\mathbf{x}_i) = \sum_{k=1}^K F_k(t|\mathbf{x}_i) = 1 - S(t|\mathbf{x}_i). \quad (4)$$

2.2 | Prognostic model development

2.2.1 | Model format for time-to-event outcomes

Since we are interested in not only predicting the risk of death from colon cancer but also the risk of death from other causes, our main approach to account for competing risks is to fit a cause-specific hazard model to each of the $k = 1, \dots, K$ events separately. Working in the cause-specific setting means that the total predicted risk is constrained to not exceed 1, which is not the case for the subdistribution modelling approach.^{26,27}

In the applied example in Section 3, we use $K = 2$ and develop one model for deaths due to colon cancer (patients who die from other causes are censored) and a second model for deaths due to other causes (patients who die due to colon cancer are censored), to mirror the approach used to develop several existing risk prediction models for cancer.

The PREDICT^{17,18} and Qcancer¹⁹ prediction models use Cox proportional hazard (PH) models. These are semi-parametric models where the baseline hazard function $h_{0k}(t)$ does not have a parametric form. However, the Breslow estimate of the baseline cumulative hazard from each model can be used to produce the cause-specific CIFs (Section 2.1). As this will give a step function, the estimates of the baseline cumulative hazard can first be smoothed if required.²³ The linear predictor, $\beta_k^T \mathbf{x}_i$, forms the parametric component of the model and can incorporate patient characteristics (\mathbf{x}_i) such as stage of tumour and age at diagnosis.^{28,29} A Cox PH model for each cause can be written as a combination of the baseline cause-specific hazard function and the corresponding covariate effects such that:

$$h_k(t|\mathbf{x}_i) = h_{0k}(t) e^{\beta_k^T \mathbf{x}_i}. \quad (5)$$

In this notation, we assume for simplicity that the same covariates \mathbf{x}_i are included in each of the cause-specific hazard models; however, this does not have to be the case.

An alternative approach is to use flexible parametric survival models (FPM). Rather than modelling on the hazard scale, these models are typically fitted on the log cumulative hazard scale and are fully parametric models.³⁰ They have the following form where a restricted cubic spline function, $\zeta_k(\ln(t)|\boldsymbol{\gamma}_k, \mathbf{k}_{0k})$, with parameters $\boldsymbol{\gamma}_k$ and knot locations \mathbf{k}_{0k} , is used to model the log cumulative baseline hazard function for cause k , $\ln[H_k(t|\mathbf{x}_i)]$ ²⁹:

$$\ln[H_k(t|\mathbf{x}_i)] = \zeta_k(\ln(t)|\boldsymbol{\gamma}_k, \mathbf{k}_{0k}) + \beta_k^T \mathbf{x}_i. \quad (6)$$

FPMs are used to illustrate the approach in Section 3, however, as shown in our previous work, the methods outlined in this paper can also be applied to Cox PH models.⁹

Whichever modelling approach is taken, penalty terms could be added to the likelihood function to help address any overfitting.³¹ This would apply in the first step of temporal recalibration when the predictor effects are estimated (see Section 2.2.4). For the period analysis approach (see Section 2.2.3), including a penalty term may help with overfitting when reducing the window size as the predictor effects are estimated on this smaller subsample but we do not consider this further here.

2.2.2 | Standard approach (not accounting for trends in survival)

The standard approach used to develop risk prediction models does not take account of any trends in survival that occur within the development dataset. Therefore, all individuals contribute toward the estimation of both the predictor effects and the baseline hazard for each of the k models regardless of when they were diagnosed. This approach can lead to over-estimating the risk for new patients if survival improves over this time. This is due to the higher mortality rate amongst the earliest diagnosed patients having a large influence on the overall cause-specific hazard functions which cancels out some of the recent improvements in survival.

2.2.3 | Period analysis to account for trends in survival

Period analysis is a technique which is often used in cancer survival to produce more up-to-date estimates of survival by limiting the use of older data where possible.^{11,32} As shown in Figure 1, a period window is defined and only the follow-up time and events that occur during the window are included in the analysis by using delayed entry techniques.^{10,33}

The hazard rates at earlier time points are only estimated from patients diagnosed during or shortly before the period window (eg, Patient E) and are therefore more up-to-date since data from earlier diagnosed patients who experienced poorer survival are excluded (eg, Patients A and B). As patients diagnosed within the window have limited follow-up available, estimates of the hazard rates at later time points cannot be estimated from this group alone. Therefore, delayed entry is used in order to include some information from earlier diagnosed patients. For instance, Patient A contributes toward the estimation of the hazard rates between 8 and 9 years but not at earlier time points since their follow-up time is left truncated.

Due to maximising the use of more recent data, period analysis has been shown to produce more accurate survival estimates for new patients.³³⁻³⁶ However, as only the events that occur during the window contribute toward the analysis, this method leads to a reduction in the sample size and number of events for model development. This may be problematic in small datasets and could lead to overfitting.³⁷

The choice of the window size is a bias-variance trade-off where a narrower window has the potential to produce more up-to-date survival estimates but results in a greater reduction in the number of events and sample size as shown in Table 2. Previous studies using period analysis have used a window between 1 and 5 years.^{35,36,38-41} In the applied example in Section 3, we use a window of 3 years and provide a sensitivity analysis in Appendices A.2 and A.4 using a range of window sizes to demonstrate how this choice impacts both the survival estimates and their uncertainty.

Patient	Year of Diagnosis and Follow-Up									
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
A	1	2	3	4	5	6	7	8	9	
B			1	2	3					
C					1	2	3	4	5	6
D							1	2	3	
E										1

FIGURE 1 Contribution of follow-up time from five hypothetical patients to a period analysis using a 3-year window of 2002–2004.

Applying period analysis when modelling on the subdistribution hazard is possible and we include an example of this in Appendix A.7. However, we would urge caution since many of the standard packages used to fit these models do not appropriately account for delayed entry.⁴² An alternative to using these packages is to first expand the data to calculate the time-dependent weights, see Appendix A.1.3 for example Stata code.

2.2.4 | Temporal recalibration to account for trends in survival

Temporal recalibration is a two-step process where the risk prediction model is first developed as usual using the full dataset to estimate the predictor effects. The model is then recalibrated by re-estimating the baseline hazard in the period analysis window whilst constraining the predictor effects to remain the same as in the original model.⁹ This maximises the use of data as the predictor effects are estimated on the full dataset in contrast to using period analysis to develop the model where the predictor effects are estimated using data within the period window.

When developing K cause-specific hazard models to account for competing risks, this process can be repeated on each of the models separately to ensure that the cause-specific survival and hazard functions are as up-to-date as possible. The risk predictions of interest can then be obtained using the equations shown in Section 2.1.

Alternatively, if competing risks were accounted for by using the approach proposed by Fine and Gray,²⁰ temporal recalibration could be applied by first fitting the standard model and then re-estimating the baseline subdistribution hazard in the period analysis subsample whilst constraining the subdistribution hazard ratios to remain the same. Changes in the other cause mortality may affect the subdistribution hazard ratios relating to the cancer mortality (and vice versa)⁴³ and hence constraining them to remain the same would be a stronger assumption than in the cause-specific setting.

2.3 | Assessing the calibration of risk predictions

Assessing the calibration of a prognostic model involves comparing the predicted risks to the observed risks. This section outlines approaches to estimate the observed risk, calculate calibration-in-the-large and produce calibration plots.

2.3.1 | Nonparametric estimator of the observed risk

The Aalen–Johansen estimator, $\hat{F}_k^{AJ}(t)$, is a nonparametric estimator of the cause-specific CIF

$$\hat{F}_k^{AJ}(t) = \sum_{j|t_j \leq t} \hat{S}^{KM}(t_{j-1}) \frac{d_{kj}}{n_j}, \quad (7)$$

where t_j is the j^{th} -ordered event time and $\hat{S}^{KM}(t_{j-1})$ is the all-cause Kaplan–Meier estimate at the previous event time.^{44,45} The cause-specific hazard rate for cause k is given by $\frac{d_{kj}}{n_j}$, where d_{kj} is the number of deaths due to cause k at time t_j and n_j is the number of individuals at risk at time t_j .¹⁵

A nonparametric estimate of the all-cause risk of death, $\hat{F}_{all}^{KM}(t)$, can be obtained using the all-cause Kaplan–Meier estimate of risk $(1 - \hat{S}^{KM}(t))$.^{16,25} These observed risk estimates then need to be compared with the model's predicted risks, to check if they calibrate well, as described below.

2.3.2 | Calibration-in-the-large

In an external validation of a model, the marginal predicted cause-specific CIFs, $\bar{F}_k(t)$, can be calculated as the average predicted risk across all N individuals:

$$\bar{F}_k(t) = \frac{1}{N} \sum_{i=1}^N \hat{F}_k(t|\mathbf{x}_i). \quad (8)$$

The Aalen-Johansen estimator can be used to quantify the observed risk and if the model is well-calibrated, these estimates should agree closely.

The marginal predicted all-cause CIF, $\bar{F}_{all}(t)$, can be calculated as the sum of the K marginal cause-specific CIFs, $\bar{F}_k(t)$:

$$\bar{F}_{all}(t) = \sum_{k=1}^K \bar{F}_k(t). \quad (9)$$

The calibration of the all-cause risk predictions can then be assessed by comparing this to the all-cause Kaplan–Meier estimate of risk.

The area under the CIFs relate to the restricted life years lost due to that cause.⁴⁶ For example, the area under the 10-year cause-specific CIF for colon cancer is the restricted life years lost up to 10 years due to colon cancer. This provides an additional measure that can be used to check calibration. If the model is well-calibrated, the area under the predicted marginal cause-specific CIF ($\bar{F}_k(t)$) should be in good agreement with the area under the Aalen-Johansen estimator. Likewise, the area under the marginal predicted all-cause CIF ($\bar{F}_{all}(t)$) should be similar to the area under the all-cause Kaplan-Meier estimate of risk.

2.3.3 | Calibration plots

Calibration-in-the-large focuses on the overall agreement between the observed and predicted risks, but it is also important to check calibration across the entire range of predictions.^{37,47} Calibration plots with calibration curves can be used to do this at each time-point of interest. For competing risks, calibration plots can be produced for the risk of death due to each cause as well as the all-cause (total) risk of death.

To avoid grouping individuals, pseudo-values can be used. A pseudo-value, $\hat{\Theta}_i$, is calculated for each individual using

$$\hat{\Theta}_i = n\hat{\Theta} - (n-1)\hat{\Theta}_{-i}, \quad (10)$$

where $\hat{\Theta}$ is the Aalen-Johansen estimator when estimating cause-specific CIFs or the all-cause Kaplan-Meier when estimating all-cause survival.⁴⁸ For the latter, $1 - \hat{\Theta}_i$ gives pseudo-values for the all-cause risk of death. The estimator is calculated using the entire cohort ($\hat{\Theta}$) and also when excluding individual i ($\hat{\Theta}_{-i}$) in order to calculate the pseudo-value for individual i ($\hat{\Theta}_i$).

Once the predicted risks and pseudo-values have been calculated for each individual, it is possible to estimate a flexible calibration curve either parametrically, for example using restricted cubic splines,⁴⁹ or non-parametrically with a method such as the nearest neighbour smoothing.⁵⁰ Here we use the latter by smoothing the pseudo-values as a function of the predicted risks.

Whilst the smooth calibration plots allow calibration to be assessed visually at a particular time point of interest t , the Integrated Calibration Index (ICI) can help to quantify how well the predictions are calibrated.^{49,51} The ICI is the mean absolute difference between the observed risks (obtained using the smoothed calibration curve) and the predicted risks (obtained from the prognostic model) across all individuals, where a value of zero would indicate perfect calibration. It can be calculated as follows,

$$ICI_t = \frac{1}{N} \sum \left| \hat{P}_t^C - \hat{P}_t \right| \quad (11)$$

where an individual's predicted risk at time t , \hat{P}_t , is obtained from the prognostic model and the value that the smoothed calibration curve takes at \hat{P}_t gives their observed risk, \hat{P}_t^C .⁵¹ The mean of the absolute difference of \hat{P}_t and \hat{P}_t^C across all N individuals gives the ICI.

2.4 | Concordance

As discussed in our original article, performing temporal recalibration on a proportional hazards model (in the non-competing risks setting) does not affect Harrell's c-index as the predictor effects are constrained to remain the same

and therefore the ordering of the participants does not change.⁹ In the competing risks setting, Wolber's concordance index could be calculated which orders individuals based on their predicted CIF for the event of interest.^{47,52} If Fine and Gray models are used with no time-dependent effects, then the concordance index will be the same for the standard approach and temporal recalibration as the predictor effects are the same.⁴⁷ However in the cause-specific setting, the CIF is dependent on both cause-specific models and therefore updating the baseline hazard of both models may result in some small changes to the ordering. Whilst we would expect the concordance index to be very similar for these methods, they will not necessarily be the same.

3 | EXAMPLE

3.1 | Data

We compare the three approaches (standard method – not accounting for trends in survival over time, temporal recalibration, period analysis) using an example of survival following a diagnosis of colon cancer using data from the Surveillance, Epidemiology and End Results (SEER) program which includes cancer registry data from nine registries in the United States.⁵³

Model development included white and black patients diagnosed aged 18–99 with colon cancer (ICD10 codes C18.0–18.9) between 1995 and 2004 (follow-up until December 31, 2004). Any duplicate records or patients with an unknown survival time or cause of death were excluded, as were any patients who had an incomplete date of diagnosis or death.

In our previous article, we used all available data to develop a prognostic model. However, as many prognostic models are not developed using large national databases, in this example, we restrict the analysis to a random 10% sample which leaves 4683 patients for model development. This allows us to highlight the impact of key modelling choices when working with smaller datasets—for example, this sample size allows us to demonstrate how the choice of period window to use with temporal recalibration or period analysis can affect the calibration and uncertainty of the risk predictions.

For simplicity, a complete case analysis was performed. However, these methods can be applied on multiply imputed data where the imputation models are adapted accordingly for competing risks.⁵⁴ Rubin's rules can then be used to combine the estimates of the predictor effects and produce the risk predictions.⁵⁵

To assess the calibration of the predictions for more recent patients, a validation dataset of all patients diagnosed in 2005 (follow-up until December 31, 2015) was used. Table 1 presents the baseline characteristics of the development ($N = 4683$) and validation ($N = 5504$) datasets after the missing data were removed and the random 10% sample for model development was selected.

3.2 | Methods

Three different strategies for model development were applied to both the colon cancer and other cause models: the standard approach (not accounting for trends in survival), temporal recalibration and period analysis.

Flexible parametric survival models were used to develop each of the cause-specific hazard models using 5 degrees of freedom (6 knots) for the log cumulative baseline hazard. Using 5 degrees of freedom was thought to provide sufficient flexibility whilst not over-parameterising the baseline. Previous simulation studies have shown flexible parametric survival models to be fairly insensitive to the number of knots used.^{56,57}

The colon cancer model included the following predictors and assumed proportional hazards: age (a non-linear effect modelled with restricted cubic splines and 3 degrees of freedom) and the categorical variables of sex, ethnicity, stage and grade of tumour. All predictors except tumour grade were included in the other cause model. Stage of tumour at diagnosis would not normally be expected to impact the other cause mortality, however it was included since the effect for Stage 3 was large (hazard ratio = 2.04, see Table 3 in Appendix A.3). The magnitude of the hazard ratio was likely due to a number of patients dying due to other causes very shortly after being diagnosed with a Stage 3 tumour. These could be incidental diagnoses of cancer when patients were seriously ill in hospital and being treated for other conditions. Due to the large hazard ratio, stage at diagnosis was included in the other cause model.

TABLE 1 Baseline characteristics of the development and validation datasets.

Variable	Mean (S.D) or N (%)	
	Development diagnosed: 1995–2004 follow-up: until 31/12/2004	Validation diagnosed: 2005 follow-up: until 31/12/2015
Age	70.1 (12.8)	69.3 (13.5)
Sex	Male	2283 (48.8%)
	Female	2400 (51.3%)
Ethnicity	White	4040 (86.3%)
	Black	643 (13.7%)
Stage	1: Localized	1725 (36.8%)
	2: Regionalized	2099 (44.8%)
	3: Distant	859 (18.3%)
Grade	1: Well differentiated	542 (11.6%)
	2: Moderately differentiated	3197 (68.3%)
	3: Poorly differentiated	900 (19.2%)
	4: Undifferentiated	44 (0.9%)
Total	4683 (100.0%)	5504 (100.0%)

TABLE 2 Sample size and number of events for each cause when using the standard approach or period analysis for model development.

Method	Full sample			Diagnosed within the window		
	Sample size	Deaths due to colon cancer	Deaths due to other causes	Sample size	Deaths due to colon cancer	Deaths due to other causes
Standard approach	4683	1176	872	4683	1176	872
Period analysis: 5-year window	3881	660	588	2419	440	262
Period analysis: 4-year window	3659	531	496	1997	325	186
Period analysis: 3-year window	3405	399	377	1505	205	113
Period analysis: 2-year window	3167	273	266	1066	126	72
Period analysis: 1-year window	2888	134	128	567	45	22

Table 2 displays the sample size and number of events for developing each of the cause-specific hazard models using the standard approach or period analysis. Although the sample size remains large even when using a 1 year period window, it includes many patients diagnosed before the window who only contribute toward the estimation of the hazard rates at later time points. Therefore, it is important to also consider the number of patients diagnosed within the window since these individuals are considered as being at risk from time zero and are the only patients to contribute toward the estimation of the hazard rates at early time points.

The use of period analysis also has a large impact on the number of events which is particularly evident in smaller datasets where only a small number of events remain in the analysis when using a narrow window. For example, when developing the models using a 3 year window, only 34% and 43% of the events were retained for the colon cancer and other cause models respectively. Although using a narrower window has the potential to produce more up-to-date predictions, it can also limit the number of predictor parameters that can be included due to the potential of overfitting. An advantage of temporal recalibration is that the predictor effects are estimated using the standard approach on the full dataset and only the baseline is re-estimated using this subsample. In Section 3.4, the results using a 3 year window are presented; however, Appendix A.4 provides a sensitivity analysis when using different window sizes.

For each of the period windows, the number of events for each of the causes was similar (see Table 2), and therefore the same window width was used to develop each of the models. However, if there were a scenario where one cause of

death was much more likely, for example lung cancer, where the majority of the deaths are due to the cancer, different windows could be used for each model to avoid overfitting in the model with fewer events. In the lung cancer example in Appendix A.5, a 2-year window was used for the lung cancer model and a 4 year window was used for the other cause mortality.

To assess the calibration of the risk predictions for more recently diagnosed patients, the marginal predicted 10-year cause-specific and all-cause CIFs were compared to the nonparametric equivalents in the validation dataset. The predicted and observed restricted life years lost to cancer and other causes up to 5 and 10 years were also calculated. To further assess calibration, calibration plots at 10 years were produced and the ICI for each plot was calculated.

3.3 | Software

The analysis was performed in Stata 17 using several user-written Stata packages: *stpm2* for fitting flexible parametric survival models,³⁰ *standsurv* for producing the risk predictions,⁵⁸ *stcompct* for calculating the Aalen-Johansen estimates,⁵⁹ *stpsurv* and *stpci* for calculating pseudo-values.⁴⁸ Example Stata code for fitting the models is provided in Appendix A.1.

3.4 | Results

3.4.1 | Choice of period window

As shown in Table 2, using a period analysis approach to develop prognostic models reduces the sample size and number of events which increase the uncertainty of the predictions and may result in overfitting in some cases. Sample size criteria⁶⁰ could be used to inform the choice of period window by calculating the minimum number of events required for a particular application and ensuring that the minimum sample size required are diagnosed within the period window.

The impact of the period window can also be informally assessed using the global shrinkage factor to determine whether there is evidence of overfitting during model development.³⁷ Using the standard approach to develop the models on all available data resulted in minimal overfitting with a global shrinkage factor of at least 0.98 for each of the models.

In practice, the predictor effects from the models developed using the standard approach should be multiplied by their corresponding uniform shrinkage factor and then constrained at their shrunken values when re-estimating the (log cumulative) baseline hazard to ensure calibration-in-the-large.³⁷ For temporal recalibration, the same shrunken values for the predictor effects can be used but the difference is that the baseline hazard would be re-estimated in the period window. However, it makes little difference in this example, as the shrinkage factors are very close to 1.

In contrast, when using period analysis, the predictor effects are estimated on the subsample of data defined by the window. Whilst the global shrinkage factor was at least 0.97 when using a 3 year window, using increasingly narrow windows led to a greater degree of overfitting, see Table 3. For simplicity here, no adjustment for overfitting was made since the shrinkage factors for all the models included in the main analysis are close to 1 but in principle the predictor effects should be constrained at the shrunken values and the baseline re-estimated. This is most important in small sample sizes where the shrinkage factor becomes far from 1.

TABLE 3 Global shrinkage factor for each of the cause-specific models.

Method	Colon cancer	Other causes
Standard approach (or temporal recalibration)	0.99	0.98
Period analysis: 5-year window	0.99	0.98
Period analysis: 4-year window	0.99	0.98
Period analysis: 3-year window	0.98	0.97
Period analysis: 2-year window	0.97	0.95
Period analysis: 1-year window	0.93	0.90

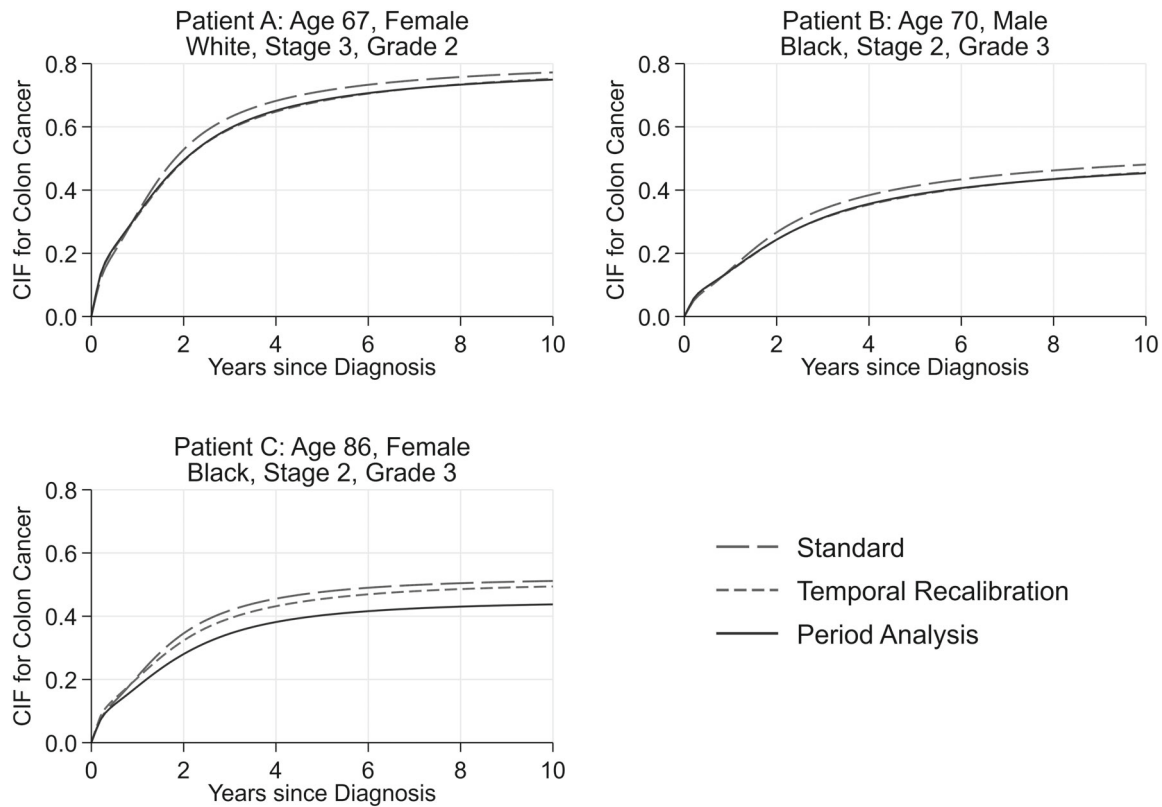


FIGURE 2 Comparison of the predicted risk of death due to colon cancer of three patients with different covariate patterns. The predictions for temporal recalibration and period analysis are from the models which used a 3-year window. The predictions from temporal recalibration and period analysis overlay almost exactly for Patients A and B.

In addition, the standard error of the predictor effects can also be examined. For example, the standard errors of the log hazard ratios are approximately twice as large when using a period analysis window of 3 years in comparison to the standard approach, see Tables 2 and 3 in Appendix A.3.

In addition to minimising overfitting, it is also important to consider the uncertainty in estimating the baseline of the model.⁶⁰ One approach to assess this could be to fit a model without any predictors and estimate the cause-specific survival functions with a 95% confidence interval to gain an understanding of the impact that using a smaller window may have due to there being fewer events. Given that temporal recalibration only estimates the baseline and keeps the covariates fixed it gives an informal guide to the impact of the choice of window size. Figures 1 and 2 in Appendix A.2 show the impact of changing the window on the width of these intervals, and hence the uncertainty of these estimates.

The choice of period window could then be guided based on a combination of the global shrinkage factor, the uncertainty in the predictor effects and the uncertainty in the baseline. In temporal recalibration, only the baseline is estimated in the period analysis subsample and therefore the most important consideration would be the precision for estimating the baseline since the predictor effects are estimated on the full dataset.

3.4.2 | Calibration

Accounting for trends in survival during model development can lead to updated risk predictions for individuals. Figure 2 displays the predicted risk of death due to colon cancer for three patients with varying characteristics. As can be seen, using temporal recalibration or period analysis produced predictions that were around 3 percentage points lower than using the standard approach for Patients A and B. For many covariate patterns, the risk predictions from temporal recalibration and period analysis were similar; however, for Patient C and other more elderly patients, using period analysis produced the lowest risk predictions that were around 7 percentage points lower than the standard method.

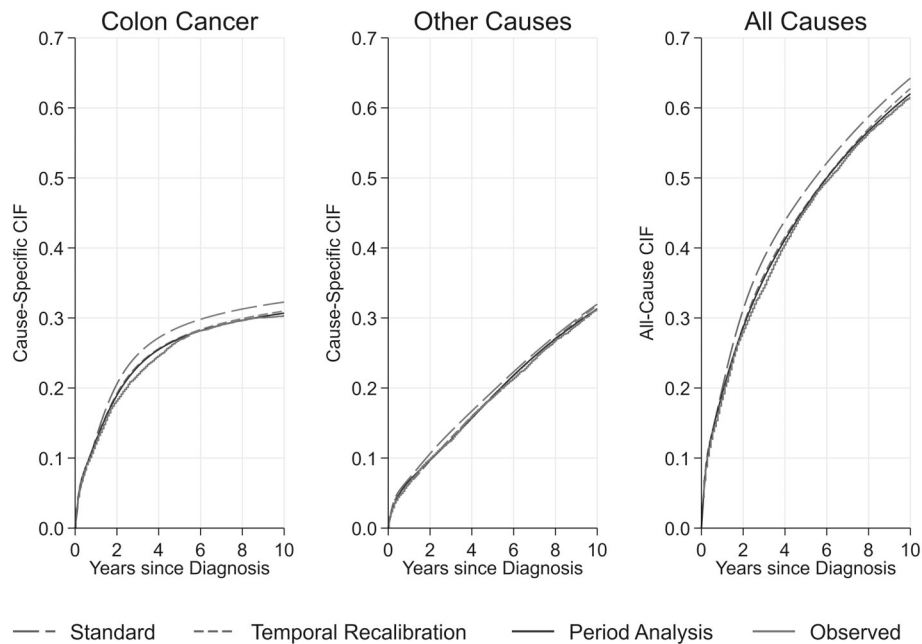


FIGURE 3 Comparison of the predicted and observed marginal cause-specific CIFs. The predictions for temporal recalibration and period analysis overlay almost exactly and are from the models which used a 3-year window.

TABLE 4 Difference between the predicted and observed CIFs at 5 and 10 years post diagnosis with a 95% confidence interval.

Time Point	Method	Colon Cancer	Other Causes	All Causes
5 years	Standard approach	0.019 [0.007, 0.030]	0.009 [−0.001, 0.019]	0.028 [0.014, 0.041]
	Temporal recalibration	0.003 [−0.008, 0.015]	0.003 [−0.008, 0.013]	0.006 [−0.007, 0.019]
	Period analysis	0.002 [−0.010, 0.014]	0.002 [−0.009, 0.012]	0.004 [−0.009, 0.017]
10 years	Standard approach	0.019 [0.007, 0.032]	0.008 [−0.004, 0.020]	0.028 [0.015, 0.041]
	Temporal recalibration	0.007 [−0.006, 0.019]	0.006 [−0.006, 0.018]	0.013 [0.000, 0.026]
	Period analysis	0.004 [−0.009, 0.016]	0.002 [−0.011, 0.014]	0.005 [−0.008, 0.018]

Note: The predictions from using a 3 year window are presented for temporal recalibration and period analysis.

Figure 3 compares the predicted marginal (average) CIFs for the validation dataset (patients diagnosed in 2005) to the observed nonparametric estimates. Using the standard approach to develop each model led to an over-estimation of 0.019, 0.008 and 0.028 for the 10-year marginal CIFs for colon cancer, other causes and all causes respectively, see Table 4. Using temporal recalibration or period analysis improved the calibration of the predicted risks across the full range of follow-up, where particular improvements can be seen in the predicted risk of death due to colon cancer and the total risk of death. Very similar results were found when modelling on the subdistribution hazard scale as shown in Appendix A.7.

The predictions at 5 years are slightly better calibrated than at 10 years for the models developed using temporal recalibration or period analysis. The predictions at earlier time points are likely to always be better calibrated since it is possible to estimate them using more recent data. For example, the 10 year hazard rates can only be estimated from patients diagnosed in 1995 since these are the only individuals who have sufficient follow-up time. In contrast, the 5 year hazard rates are estimated from more recently diagnosed patients (diagnosed between 1997 and 2000) in the temporal recalibration and period approach and the estimates are therefore more up-to-date.

Improvements in calibration can also be seen in the calibration plots in Figure 4, where using either temporal recalibration or period analysis led to a reduction in the Integrated Calibration Index. For all methods, the risk predictions for death due to other causes under-estimated the risk in the low-risk patients. However, this only affected a small number of patients as only 49 had a cause-specific CIF less than 0.05.

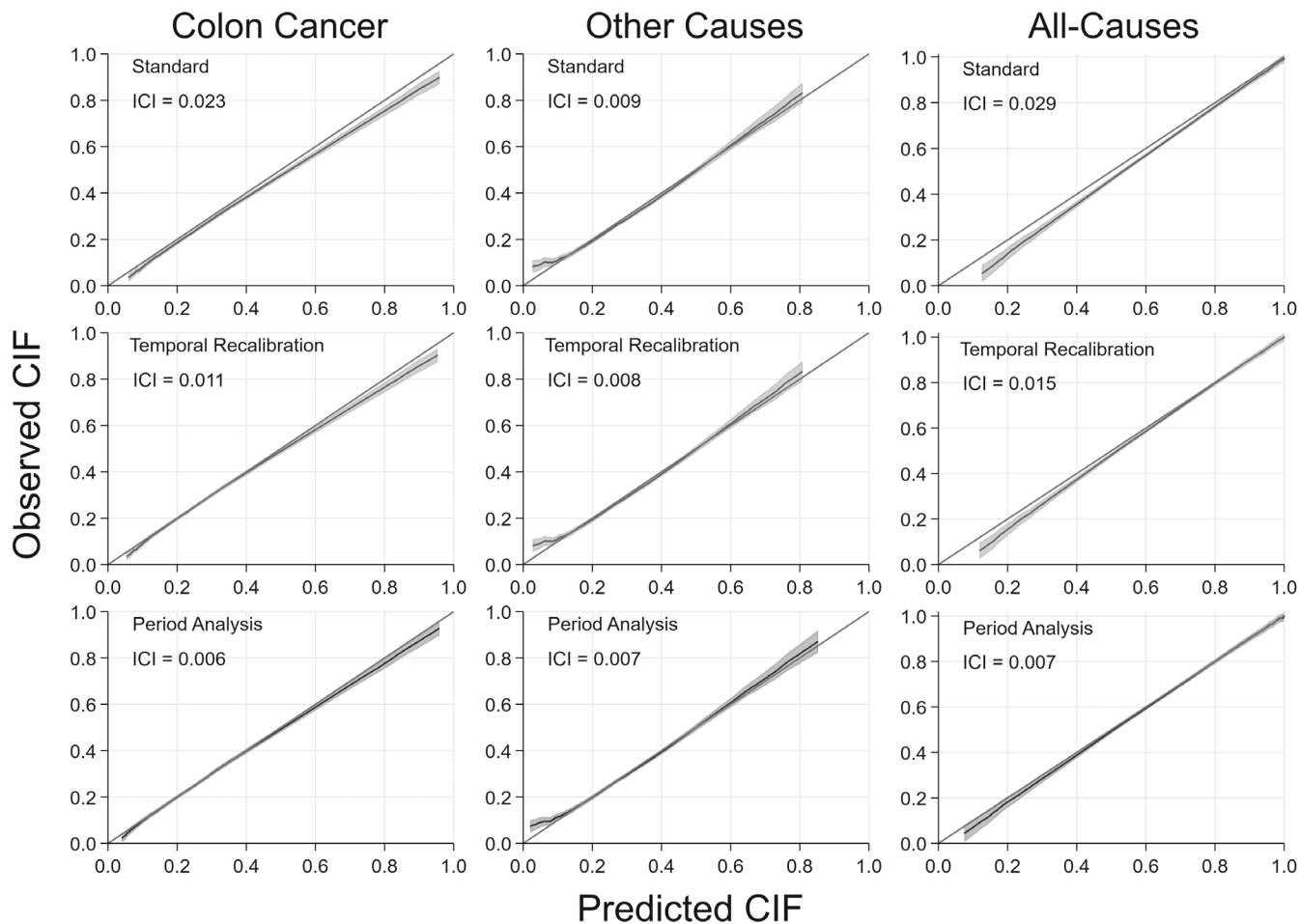


FIGURE 4 Calibration plots at 10 years for each of the cause-specific CIFs and the total risk. The results from using a 3-year window are presented for the temporal recalibration and period analysis models.

The marginal CIFs using different window sizes for temporal recalibration and period analysis are presented in Figures 3–6 in Appendix A.4. Whilst using a 5 year window led to greater precision, a 2 year window further improved the calibration of the predictions. This demonstrates the bias-variance trade-off where better calibrated risk predictions can be produced using a narrower window at the cost of greater uncertainty. When selecting the window, it is important to ensure that there are a sufficient number of events to estimate the baseline (and the predictor effects if using period analysis) reliably.

The area under the CIFs in Figure 3 relates to the restricted life years lost up to 10 years.⁴⁶ By improving the calibration of the risk predictions, the estimates of the life years lost due to colon cancer and the total life years lost agreed more closely with the observed nonparametric estimators as shown in Table 5. It is important to account for trends in survival in each of the cause-specific models. If only one of the models were temporally recalibrated, the other model would remain miscalibrated and although the calibration of the total risk predictions would improve, the proportion of life years lost to each cause would be incorrect.

Although the performance of temporal recalibration and period analysis was similar, differences in risk predictions were identified for certain covariate patterns, in particular for the most elderly patients where period analysis produced much lower risk predictions for death due to colon cancer (eg, Patient C, Figure 2). This was further investigated by producing calibration plots that only included those aged 85 and over at diagnosis, see Figure 5. Here, it can be seen that using period analysis under-estimated the CIF whilst using temporal recalibration produced better calibrated risk predictions. These differences may be due to the greater uncertainty at which the hazard ratios for age can be estimated in period analysis, particularly when there is sparser data in the upper tail of the age distribution. For example, there were 532 patients aged 85 and over in the model development dataset. However, this reduced to 315 when using period

TABLE 5 Restricted life years lost up to 5 and 10 years and the restricted life years lost due to colon cancer for the validation dataset.

Time Point	Method	Restricted life years lost	Restricted life years lost due to cancer (%)
5 years	Aalen-Johansen	1.47	0.92 (62.6%)
	Standard approach	1.61	1.02 (63.2%)
	Temporal recalibration	1.52	0.97 (63.6%)
	Period analysis	1.51	0.96 (63.5%)
10 years	Aalen-Johansen	4.21	2.39 (56.7%)
	Standard approach	4.46	2.56 (57.4%)
	Temporal recalibration	4.27	2.44 (57.0%)
	Period analysis	4.25	2.42 (57.0%)

Note: The results for temporal recalibration and period analysis were from using a 3-year window.

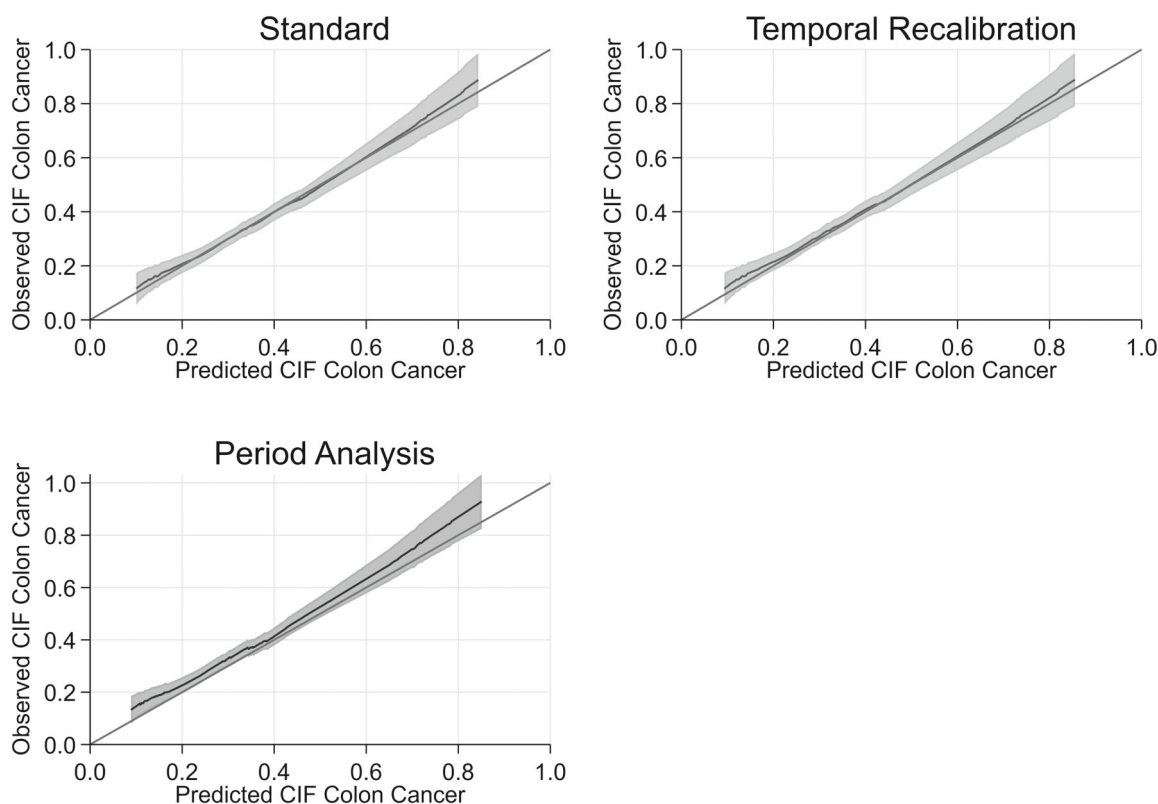


FIGURE 5 Calibration plots at 10 years for the cause-specific CIF due to colon cancer for patients aged 85 and over at diagnosis. The results from using a 3-year window are presented for the temporal recalibration and period analysis models.

analysis with a 3 year window since many of these older patients have short survival times and therefore did not survive into the window to be included in the analysis.

4 | DISCUSSION

Not accounting for trends in survival over calendar time when developing prognostic models can lead to miscalibrated risk predictions for new patients. In this example, not accounting for the improvements in survival following a diagnosis

of cancer resulted in over-estimating the risk and consequently under-estimating the long-term survival when the model was temporally validated in the same population but with more recently diagnosed patients. Using period analysis can lead to better calibrated predictions but results in a reduction in the sample size and number of events that can be used to develop the model.

Temporal recalibration can also be used to produce more up-to-date risk predictions by first estimating the predictor effects on the full dataset and then re-estimating the baseline hazard using delayed entry techniques to capture improvements in survival. Due to the lack of long-term follow-up data for the most recently diagnosed patients, standard recalibration techniques can often only be applied once new data are available. For example, the most recently diagnosed patients in the development dataset may have a maximum of 3 years of follow-up and therefore it would not be possible to update the 10-year risk estimates without additional follow-up information. In contrast, temporal recalibration simply uses a subset of the original data to update the model parameters. This allows improvements in survival to be accounted for at the model development stage. This method can be extended into a competing risk setting by repeating this process for each of the cause-specific hazard models separately and then using these more up-to-date hazard and survival functions to produce the risk predictions.

In the applied example for colon cancer survival, we showed that using either period analysis or temporal recalibration improved the calibration-in-the-large and the calibration within different risk groups both for the risk due to colon cancer and the total risk of death. In this case, the development data only spanned 10 years, however, if older data were also included, or there were larger improvements in survival over this time period, using delayed entry methods would have a larger impact.

In datasets with a large sample size and number of events, period analysis is likely to perform very well. In addition to providing an updated baseline, more up-to-date predictor effects can be obtained if there have been substantial changes over time, since they are estimated on a recent subsample of data. However, overfitting will be more problematic with period analysis in situations with small sample sizes. In this particular example, although there were only 399 and 377 events in each cause-specific hazard model when a 3-year period window was used, the prognostic model was relatively simplistic and only included a small number of predictor parameters with no interaction terms or time-dependent effects. Therefore, the amount of shrinkage remained small despite the number of events being reduced by around 60% when using period analysis. If a more complex model were to be developed using this dataset, the issue of overfitting would be more problematic when using period analysis in comparison to estimating the predictor effects on the full dataset.

Overfitting would also be an issue when using period analysis in small datasets and in these settings, shrinkage and penalisation methods may not always be reliable.⁶¹ Developing prognostic models in competing risk settings can be particularly complex since a sufficient number of events must be available to develop each of the cause-specific hazard models in order to produce well-calibrated risk predictions. Therefore, even if the event of interest is common, the competing events may occur less frequently and hence overfitting may be present in at least one of the models. An example of this could be lung cancer, where there were approximately four times as many deaths due to cancer than other causes (see Appendix A.5).

In contrast, temporal recalibration utilises all available patient data to estimate the predictor effects which will likely make this method more stable in these settings. As the predictor effects can be estimated more precisely, this method is particularly useful when there are rare covariate patterns. For example, temporal recalibration produced better calibrated risk predictions for patients aged over 85 at diagnosis. Therefore, whilst period analysis will likely perform very well in large datasets where overfitting is not an issue, temporal recalibration is a more appropriate method in smaller datasets. With either method, careful thought is required when selecting an appropriate period window to ensure that there are a sufficient number of events to estimate the model parameters reliably.

In our main analysis we focused on modelling in the cause-specific setting but we also demonstrate in Appendix A.7 that very similar improvements in calibration can be made when using temporal recalibration or period analysis with Fine and Gray models. Although it is possible to apply these approaches in this setting, caution must be taken as many of the standard software packages used to fit models on the subdistribution hazard scale do not appropriately account for delayed entry.⁴²

In Section 3, we only showed a simple example using FPMs to illustrate the process of fitting these types of models under a complete case analysis and assuming proportional hazards. However, these methods can be used in a range of model formats including Cox PH models, in conjunction with multiple imputation and when the PH assumption is not valid.

Recommendations

- In small datasets, temporal recalibration is our preferred approach since only the baseline is estimated using the recent subsample of data. As the predictor effects are estimated using all the data this limits overfitting in comparison to period analysis.
- In large datasets, both temporal recalibration and period analysis are likely to perform well in estimating a more up-to-date baseline hazard. Due to the size of the data there will be a large number of events even when using period analysis and so overfitting should be minimal using either method. In settings where the predictor effects also change substantially over time, using period analysis would be advantageous since the predictor effects are estimated using the more recent subsample and will therefore be more up-to-date.

Using temporal recalibration in a competing risks setting is directly applicable to existing prognostic models for cancer such as PREDICT Breast,¹⁷ PREDICT Prostate¹⁸ and QCaner Colorectal Survival,¹⁹ all of which use one cause-specific hazard model for deaths due to the cancer of interest and a second for deaths due to other causes. Using temporal recalibration could provide a suitable approach to update these types of models without the need for any additional data. However, when models are developed using databases that are regularly updated, incorporating this more recent data when performing temporal recalibration has the potential to make additional improvements in calibration.

In addition, temporal recalibration could also be used if models are to be continually updated over time when new data become available.⁶² For example, in our previous article, we demonstrated that if the predictor effects are stable over time, the baseline can simply be updated in a more recent period window in the extended dataset that includes both the original and new data.⁹ We also showed an alternative approach where the model was first re-fitted with, for example, the latest 10 years of data (using the standard approach). This allows new estimates of the predictor effects to be calculated which is advantageous if there have been any changes over time. This model can then be temporally recalibrated to update the baseline. Both of these methods extend naturally into the competing risk settings by updating each cause-specific model separately. In summary, temporal recalibration can be applied in a wide range of settings and is well-suited for developing risk prediction models with competing risks, particularly when cause-specific hazard models are used. By accounting for improvements in survival at the model development stage, better calibrated risk predictions for new patients can be produced.

CONFLICT OF INTEREST STATEMENT

Sarah Booth, Lucinda Archer, Joie Ensor, Richard D. Riley, Paul C. Lambert, Mark J. Rutherford: None. Sarwar I. Mozumder: Employed by Roche Products Ltd and AstraZeneca for work unrelated to this research during the drafting of the manuscript.

ACKNOWLEDGEMENTS

This study was supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration East Midlands (ARC EM) and the Leicester NIHR Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. We would like to thank the Associate Editor and the two reviewers for providing very useful feedback to improve our paper.

FUNDING INFORMATION

Sarah Booth was supported by Cancer Research UK project grant (C14183/A29739). This work was supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. Sarwar I. Mozumder was supported by the National Institute for Health Research (NIHR Advanced Fellowship, Dr Sarwar Mozumder, NIHR300100). Paul C. Lambert received support from the Swedish Cancer Society (Cancerfonden) (grant number 2018/744) and the Swedish


Research Council (Vetenskapsrådet) (grant number 2017-01591). Mark J. Rutherford received support from a Cancer Research UK project grant (C41379/A27583).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available through the Surveillance, Epidemiology, and End Results Program (SEER). Access to the database can be requested here: <https://seer.cancer.gov/data/access.html>.

ORCID

Sarah Booth  <https://orcid.org/0000-0003-1799-3144>

Sarwar I. Mozumder  <https://orcid.org/0000-0001-9644-7525>

Lucinda Archer  <https://orcid.org/0000-0003-2504-2613>

Joie Ensor  <https://orcid.org/0000-0001-7481-0282>

Richard D. Riley  <https://orcid.org/0000-0001-8699-0735>

Paul C. Lambert  <https://orcid.org/0000-0002-5337-663X>

Mark J. Rutherford  <https://orcid.org/0000-0003-1557-6697>

REFERENCES

1. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61(1):76-86.
2. Liou TG, Kartsonaki C, Keogh RH, Adler FR. Evaluation of a five-year predicted survival model for cystic fibrosis in later time periods. *Sci Rep*. 2020;10(1):6602.
3. Siregar S, Nieboer D, Versteegh MIM, Steyerberg EW, Takkenberg JJM. Methods for updating a risk prediction model for cardiac surgery: a statistical primer. *Interact Cardiovasc Thorac Surg*. 2019;28(3):333-338.
4. te Velde ER, Nieboer D, Lintsen AM, et al. Comparison of two models predicting IVF success: the effect of time trends on model performance. *Hum Reprod*. 2014;29(1):57-64.
5. Hickey GL, Grant SW, Murphy GJ, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur J Cardiothorac Surg*. 2013;43(6):1146-1152.
6. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. Basel, Switzerland: Springer; 2019.
7. van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
8. Hippisley-Cox J, Coupland C. QRISK2. Annual Update Information. 2016 <https://qrisk.org/2017/QRISK2-2016-Annual-Update-Information.pdf>
9. Booth S, Riley RD, Ensor J, Lambert PC, Rutherford MJ. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int J Epidemiol*. 2020;49(4):1316-1325.
10. Brenner H, Gefeller O. An alternative approach to monitoring cancer patient survival. *Cancer*. 1996;78(9):2004-2010.
11. Allemani C, Matsuda T, Di Carlo V, et al. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37,513,025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet*. 2018;391(10125):1023-1075.
12. Arnold M, Rutherford MJ, Bardot A, et al. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995-2014 (ICBP SURVMARK-2): a population-based study. *Lancet Oncol*. 2019;20(11):1493-1505.
13. Coleman M, Forman D, Bryant H, et al. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. *Lancet*. 2011;377(9760):127-138.
14. Iacobucci G. Cancer survival in England: rates improve and variation falls. *BMJ*. 2019;365:l1532.
15. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26(11):2389-2430.
16. Wolbers M, Koller MT, Stel VS, et al. Competing risks analyses: objectives and approaches. *Eur Heart J*. 2014;35(42):2936-2941.
17. Candido dos Reis FJ, Wishart GC, Dicks EM, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Res*. 2017;19(1):58.
18. Thurtle DR, Greenberg DC, Lee LS, Huang HH, Pharoah PD, Gnanapragasam VJ. Individual prognosis at diagnosis in non-metastatic prostate cancer: Development and external validation of the PREDICT prostate multivariable model. *PLoS Med*. 2019;16(3):e1002758.
19. Hippisley-Cox J, Coupland C. Development and validation of risk prediction equations to estimate survival in patients with colorectal cancer: cohort study. *BMJ*. 2017;357:j2497.
20. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94(446):496-509.
21. Hinchliffe SR, Lambert PC. Extending the flexible parametric survival model for competing risks. *Stata J*. 2013;13(2):344-355.

22. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Stat Med*. 2017;36(27):4391-4400.
23. Hinchliffe SR, Lambert PC. Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC Med Res Methodol*. 2013;13(13). <https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-13-13#citeas>
24. Kipourou DK, Charvat H, Rachtel B, Belot A. Estimation of the adjusted cause-specific cumulative probability using flexible regression models for the cause-specific hazards. *Stat Med*. 2019;38(20):3896-3910.
25. Mozumder SI, Rutherford MJ, Lambert PC. stpm2cr: A flexible parametric competing risks model using a direct likelihood approach for the cause-specific cumulative incidence function. *Stata J*. 2017;17(2):462-489.
26. Mozumder SI, Rutherford MJ, Lambert PC. Direct likelihood inference on the cause-specific cumulative incidence function: a flexible parametric regression modelling approach. *Stat Med*. 2018;37(1):82-97.
27. Austin PC, Steyerberg EW, Putter H. Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. *Stat Med*. 2021;40(19):4200-4212.
28. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B: Stat Methodol*. 1972;34(2):187-220.
29. Royston P, Lambert PC. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. College Station, TX: Stata Press; 2011.
30. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *Stata J*. 2009;9(2):265-290.
31. Moons KGM, Donders ART, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol*. 2004;57(12):1262-1270.
32. Allemani C, Weir HK, Carreira H, et al. Global surveillance of cancer survival 1995–2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet*. 2015;385(9972):977-1010.
33. Smith LK, Lambert PC, Botha JL, Jones DR. Providing more up-to-date estimates of patient survival: a comparison of standard survival analysis with period analysis using life-table methods and proportional hazards models. *J Clin Epidemiol*. 2004;57(1):14-20.
34. Talbäck M, Rosén M, Stenbeck M, Dickman PW. Cancer patient survival in Sweden at the beginning of the third millennium – predictions using period analysis. *Cancer Causes Control*. 2004;15(9):967-976.
35. Ellison LF. An Empirical Evaluation of period survival analysis using data from the Canadian Cancer Registry. *Ann Epidemiol*. 2006;16(3):191-196.
36. Houterman S, Janssen-Heijnen MLG, van de Poll-Franse LV, Brenner H, Coebergh JWW. Higher long-term cancer survival rates in south-eastern Netherlands using up-to-date period analysis. *Ann Oncol*. 2006;17(4):709-712.
37. Riley RD, van der Windt D, Croft P, Moons KGM. *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. Oxford: Oxford University Press; 2019.
38. Maringe C, Belot A, Rachtel B. Prediction of cancer survival for cohorts of patients most recently diagnosed using multi-model inference. *Stat Methods Med Res*. 2020;29(12):3605-3622.
39. Mozumder SI, Dickman PW, Rutherford MJ, Lambert PC. InterPreT cancer survival: A dynamic web interactive prediction cancer survival tool for health-care professionals and cancer epidemiologists. *Cancer Epidemiol*. 2018;56:46-52.
40. Keogh RH, Szczesniak R, Taylor-Robinson D, Bilton D. Up-to-date and projected estimates of survival for people with cystic fibrosis using baseline characteristics: A longitudinal study using UK patient registry data. *J Cyst Fibros*. 2018;17(2):218-227.
41. Brenner H, Hakulinen T. Advanced detection of time trends in long-term cancer patient survival: experience from 50 years of cancer registration in Finland. *Am J Epidemiol*. 2002;156(6):566-577.
42. Bakoyannis G, Touloumi G. Impact of dependent left truncation in semiparametric competing risks methods: A simulation study. *Commun Stat Simul Comput*. 2017;46(3):2025-2042.
43. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. *Am J Epidemiol*. 2009;170(2):244-256.
44. Geskus RB. *Competing risks: Aims and methods*. *Handbook of Statistics*. Amsterdam: Elsevier; 2020.
45. Lambert PC. The estimation and modelling of cause-specific cumulative incidence functions using time-dependent weights. *Stata J*. 2017;17(1):181-207.
46. Andersen PK. Decomposition of number of life years lost according to causes of death. *Stat Med*. 2013;32(30):5278-5285.
47. Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology*. 2009;20(4):555-561.
48. Overgaard M, Andersen PK, Parner ET. Regression analysis of censored data using pseudo-observations: an update. *Stata J*. 2015;15(3):809-821.
49. Austin PC, Harrell FE, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Stat Med*. 2020;39(21):2714-2742.
50. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Stat Med*. 2014;33(18):3191-3203.
51. Austin PC, Putter H, Giardiello D, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. *Diagn Progn Res*. 2022;6(1)2. <https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-021-00114-6#citeas>
52. Wolbers M, Blanche P, Koller MT, Witteman JCM, Gerds TA. Concordance for prognostic models with competing risks. *Biostatistics*. 2014;15(3):526-539.
53. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973–2015), National Cancer Institute, DCCPS, Surveillance Research Program, released April 2018, based on the November 2017 submission, 2017.
54. Bartlett JW, Taylor JMG. Missing covariates in competing risks analysis. *Biostatistics*. 2016;17(4):751-763.
55. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons; 1987.

56. Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *J Stat Comput Simul.* 2013;85(4):777-793.
57. Syriopoulou E, Mozumder SI, Rutherford MJ, Lambert PC. Robustness of individual and marginal model-based estimates: A sensitivity analysis of flexible parametric models. *Cancer Epidemiol.* 2019;58:17-24.
58. Lambert P, Crowther MJ. Standsurv: Stata module to compute standardized (marginal) survival and related functions, Statistical Software Components S458991, Boston College Department of Economics, 2021.
59. Coviello V, Boggess M. Cumulative incidence estimation in the presence of competing risks. *Stata J.* 2004;4(2):103-112.
60. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II – binary and time-to-event outcomes. *Stat Med.* 2019;38(7):1276-1296.
61. Riley RD, Snell KI, Martin GP, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *J Clin Epidemiol.* 2021;132:88-96.
62. Jenkins DA, Martin GP, Sperrin M, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res.* 2021;5:1.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Booth S, Mozumder SI, Archer L, et al. Using temporal recalibration to improve the calibration of risk prediction models in competing risk settings when there are trends in survival over time. *Statistics in Medicine.* 2023;1-18. doi: 10.1002/sim.9898