

Developing prediction models to estimate the risk of two survival outcomes both occurring: A comparison of techniques

Pate, Alexander; Sperrin, Matthew; Riley, Richard D.; Sergeant, Jamie C.; Van Staa, Tjeerd; Peek, Niels; Mamas, Mamas A.; Lip, Gregory Y. H.; O'Flaherty, Martin; Buchan, Iain; Martin, Glen P.

DOI:
[10.1002/sim.9771](https://doi.org/10.1002/sim.9771)

License:
Creative Commons: Attribution (CC BY)

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

Citation for published version (Harvard):
Pate, A, Sperrin, M, Riley, RD, Sergeant, JC, Van Staa, T, Peek, N, Mamas, MA, Lip, GYH, O'Flaherty, M, Buchan, I & Martin, GP 2023, 'Developing prediction models to estimate the risk of two survival outcomes both occurring: A comparison of techniques', *Statistics in Medicine*. <https://doi.org/10.1002/sim.9771>

[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

Developing prediction models to estimate the risk of two survival outcomes both occurring: A comparison of techniques

Alexander Pate¹ | Matthew Sperrin¹ | Richard D. Riley² |
Jamie C. Sergeant^{3,4} | Tjeerd Van Staa¹ | Niels Peek¹ | Mamas A. Mamas⁵ |
Gregory Y. H. Lip^{6,7} | Martin O'Flaherty⁸ | Iain Buchan⁸ | Glen P. Martin¹

¹Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK

²Institute of Applied Health Research, University of Birmingham, Birmingham, UK

³Centre for Epidemiology Versus Arthritis, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

⁴Centre for Biostatistics, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

⁵Keele Cardiovascular Research Group, Keele University, Stoke-on-Trent, UK

⁶Liverpool Centre for Cardiovascular Science at University of Liverpool, Liverpool John Moores University and Liverpool Heart & Chest Hospital, Liverpool, UK

⁷Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

⁸Institute of Population Health, Faculty of Health and Life Sciences, University of Liverpool, Liverpool, UK

Correspondence

Alexander Pate, Vaughan House, University of Manchester, Manchester, M13 9GB, UK.
Email: alexander.pate@manchester.ac.uk

Funding information

Medical Research Council, Grant/Award Number: MR/T025085/1

Introduction: This study considers the prediction of the time until two survival outcomes have both occurred. We compared a variety of analytical methods motivated by a typical clinical problem of multimorbidity prognosis.

Methods: We considered five methods: product (multiply marginal risks), dual-outcome (directly model the time until both events occur), multistate models (msm), and a range of copula and frailty models. We assessed calibration and discrimination under a variety of simulated data scenarios, varying outcome prevalence, and the amount of residual correlation. The simulation focused on model misspecification and statistical power. Using data from the Clinical Practice Research Datalink, we compared model performance when predicting the risk of cardiovascular disease and type 2 diabetes both occurring.

Results: Discrimination was similar for all methods. The product method was poorly calibrated in the presence of residual correlation. The msm and dual-outcome models were the most robust to model misspecification but suffered a drop in performance at small sample sizes due to overfitting, which the copula and frailty model were less susceptible to. The copula and frailty model's performance were highly dependent on the underlying data structure. In the clinical example, the product method was poorly calibrated when adjusting for 8 major cardiovascular risk factors.

Discussion: We recommend the dual-outcome method for predicting the risk of two survival outcomes both occurring. It was the most robust to model misspecification, although was also the most prone to overfitting. The clinical example motivates the use of the methods considered in this study.

KEYWORDS

clinical prediction model, multiple outcome, multivariate, simulation, survival analysis, time-to-event

1 | BACKGROUND

Prognostic clinical prediction models (CPMs) use information that is available about an individual to estimate the risk of a future clinical event.¹⁻³ Often, the outcome being predicted is a time-to-event, such as time-to-diagnosis of cardiovascular disease. Most CPMs developed in the literature focus on predicting only a single outcome.⁴⁻¹² However, this is a suboptimal approach when clinical action relies on the likely prognosis across multiple events/outcomes. For example, consider CHADS₂ and the CHA₂DS₂-VASc scores^{13,14} which estimates the risk of stroke in patients with atrial fibrillation and can be used to help decide whether to prescribe oral anti-coagulants. The risk of stroke is balanced against the risk of major bleeding when on anti-coagulants, which can be calculated using the HAS-BLED score.¹⁵ These risks are also non-static and alter with ageing and incident comorbidities.^{16,17} Here, one is interested in predicting the risk of the outcomes both occurring (also referred to as the risk of “both-of-two” outcomes). A second example where modeling the risk of two survival outcomes both occurring is required is the prediction of local recurrence and distant metastasis of cancer, where clinical actions can depend on these developing in isolation or together (and in which order).¹⁸⁻²¹ A third example is prediction of multimorbidity risk, which is becoming more prevalent in many countries with people living longer with more than one long-term condition,²²⁻²⁷ which is having a major impact on health systems globally.²⁸⁻³⁴ While there is a plethora of CPMs developed to predict risks of common noncommunicable diseases, these are each developed in isolation, meaning they cannot appropriately model the risk of multiple long-term conditions occurring together. Also, many comorbidities tend to cluster together, as well as alter over time with ageing and changes in risk factors, with implications for clinical outcomes such as stroke and bleeding.^{35,36} Thus, such risk estimation would allow policy makers to predict future levels of multimorbidity in the population and target resources accordingly, for example, to preventive measures for key comorbidities.

A potential reason for the lack of such risk prediction in practice is that it is currently unclear which of the available methods has the best predictive performance. The simplest approach is to develop univariate survival CPMs for each outcome separately. If there is no residual correlation between the outcomes at all-time points (after conditioning on predictor variables), it is appropriate to multiply the corresponding risk scores from the univariate models to obtain the desired risk. However, such risk estimates will be miscalibrated if the (conditional) independence assumption does not hold. This concept is formally motivated in the following section. The extent of this miscalibration and how it may impact clinical practice is not clear. Currently, no study has been undertaken to compare the methods that can model the risk of both outcomes occurring when this assumption is violated.

Therefore, the aim of this study was 2-fold: (1) to measure the extent of the miscalibration in prediction of the risk of both-of-two survival outcomes using univariate models, when there is residual correlation in the outcomes and (2) to compare the performance of a variety of methods that could be used for predicting the risk of both-of-two survival outcomes. Section 2 outlines each of the prediction approaches considered in this study. Section 3 contains a simulation comparing the performance of these methods. Section 4 is a clinical example considering the prediction of cardiovascular disease (CVD) and type 2 diabetes (T2D). Section 5 contains a discussion of the results from the simulation and clinical example, and an overall discussion and recommendations.

2 | METHODS TO PREDICT THE RISK OF TWO SURVIVAL OUTCOMES BOTH OCCURRING

This section contains a summary of each method and how they can be used to estimate the risk of two survival outcomes both occurring. Note that the focus of this study is on prediction of the risk of outcomes that do not prevent the other from happening (ie, non-competing events). The prediction of time-to-event outcomes in the presence of a competing risk has been covered extensively³⁷⁻⁴⁴ and will not be the focus of this article.

2.1 | Notation and preliminaries

Let T_A and T_B be the event times for two (non-competing) events A and B , and T_C be the time until censoring. For each individual we observe $T_{A^*} = \min\{T_A, T_C\}$ and $T_{B^*} = \min\{T_B, T_C\}$. Let δ_A be the event indicator for outcome A , such that $\delta_A = 1$ if $T_A = T_{A^*}$, otherwise $\delta_A = 0$. δ_B is defined similarly. Let X be a vector of baseline predictor variables, each of which might be predictive of A or B (or both). We assume one is primarily interested in estimating the risk of outcomes

A and B occurring before timepoint t , given X ; that is, $P(T_A \leq t, T_B \leq t|X)$. However, each of the methods can provide varying levels of insight beyond estimating this quantity, such as estimating marginal risk scores, the level of dependence between the two outcomes or the temporal ordering of events. We discuss the extra utility of each method in the discussion (Section 5.3). We assume a common censoring mechanism to both outcomes throughout this study which is likely to hold in the majority of scenarios, however all the following models can still be applied when A and B have different censoring mechanisms. We assume the censoring process is independent of A and B given X , and discuss the ability of the methods to account for informative censoring and implement competing risks analysis (with respect to a third competing event) in Section 2.7.

2.2 | The product method

The product method is the most straightforward approach. One first develops univariate models for each outcome individually. We used Cox models in this study, but any time-to-event model could be used (eg, flexible parametric survival models). Univariate models for A and B allow the estimation of the marginal survival functions, $P(T_A > t|X)$ and $P(T_B > t|X)$, and the marginal risks $P(T_A \leq t|X) = 1 - P(T_A > t|X)$ and $P(T_B \leq t|X) = 1 - P(T_B > t|X)$. Under the assumption of conditional independence of A and B given X , the product of these will be an unbiased estimator of the risk.

$$P(T_A \leq t, T_B \leq t|X) = P(T_A \leq t|X) * P(T_B \leq t|X).$$

However, as the level of residual correlation increases, miscalibration of the product method in estimating $P(T_A \leq t, T_B \leq t|X)$ will increase; we examine the extent of this in the simulation study.

2.3 | Dual-outcome approach

The second method is to re-define the outcome as being the time until both outcome events have occurred, and develop a univariate model to predict this new “dual-outcome”. Let $T_{AB} = \max\{T_A, T_B\}$, and $T_{AB^*} = \min\{T_A, T_B\}$, and $\delta_{AB} = 1$ if $T_{AB} = T_{AB^*}$, otherwise $\delta_{AB} = 0$. Then a univariate model (Cox proportional hazards model or otherwise) can be developed on AB to estimate:

$$P(T_{AB} \leq t|X) = 1 - P(T_{AB} > t|X).$$

Given that $\{T_{AB} \leq t\} \Leftrightarrow \{T_A \leq t, T_B \leq t\}$, then:

$$P(T_A \leq t, T_B \leq t|X) = P(T_{AB} \leq t|X).$$

Therefore, the dual-outcome approach can provide estimates of $P(T_A \leq t, T_B \leq t|X)$ but does not have the ability to calculate marginal risk scores for each outcome in isolation.

2.4 | Copulas

Copulas are implemented by defining a dependence structure between two marginal cumulative distribution functions. The general framework is not restricted to survival models, but they have garnered a lot of attention in this area. For two survival outcomes, A and B , the survival function for both outcomes is defined as:

$$P(T_A > t, T_B > t|X) = C_\theta \{P(T_A > t|X), P(T_B > t|X)\},$$

where C_θ is the bivariate copula, a function of a parameter θ that represents the degree of dependence between A and B . After a given copula has been chosen, the parameter θ is estimated by either estimating the parameters from the marginal distributions, and then estimating the copula parameter(s) (the two-step approach),⁴⁵ or a joint likelihood can be maximized to estimate the marginal likelihood parameters and the copula parameter simultaneously.⁴⁶⁻⁴⁸

Some common examples of bivariate copulas (as given by Emura et al⁴⁹) are:

The independence copula:

$$C(u, v) = uv.$$

The Clayton copula:

$$C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, \theta > 0.$$

The Gumbel copula:

$$C_\theta(u, v) = \exp \left[- \left\{ (-\log(u))^{\theta+1} + (-\log(v))^{\theta+1} \right\}^{\frac{1}{\theta+1}} \right], \theta \geq 0.$$

The Farlie-Gumbel-Morgenstern (FGM) copula:

$$C_\theta(u, v) = uv\{1 + \theta(1-u)(1-v)\}, -1 \leq \theta \leq 1.$$

An explanation of copulas for multivariate survival analysis is given by Georges et al,⁵⁰ as well as concise summaries by Govindarajulu and D'Agostino,⁴³ and comprehensively covered in the books by Nelsen⁵¹ and Emura et al.⁴⁹ Note that these references are concerned with modeling the survival function $P(T_A > t, T_B > t|X) = C_\theta \{P(T_A > t|X), P(T_B > t|X)\}$, whereas we are interested in estimating the risk $P(T_A \leq t, T_B \leq t|X)$. This risk can be estimated using the following equation:

$$P(T_A \leq t, T_B \leq t|X) = 1 - P(T_A > t|X) - P(T_B > t|X) + C_\theta \{P(T_A > t|X), P(T_B > t|X)\}.$$

A simple proof of this equation is given in supporting information file 1.

An advantage of using copulas for risk prediction is that they explicitly model the association between the outcomes, providing a very clear framework in which to model the dependence between the outcomes. A potential drawback is that a parametric correlation structure (the copula itself) must be assumed, meaning the results may be sensitive to the choice of copula. This is something we explore through the simulation in Section 3. Note that implementing the independence copula would be analogous to the product method. To fit the copula models we implemented the joint estimation approach of Marra et al,⁴⁸ implemented in the package GJRM.^{52,53}

2.5 | Frailty models

A frailty model is a survival model with a random effect term to account for unexplained heterogeneity in survival times.⁵⁴⁻⁵⁶ Shared frailty models are generally applied to data which has a multilevel structure.⁵⁷ We propose that shared frailty models could be used to model the dependence between two outcomes, and subsequently the risk of them both occurring. To do this using a Cox framework, the following model would be fit (introducing subscript i to denote individual i):

$$h_A(t|\omega_i) = \omega_i h_{0,A}(t) \exp(\beta_A X_i),$$

$$h_B(t|\omega_i) = \omega_i h_{0,B}(t) \exp(\beta_B X_i),$$

where ω_i is the shared random effect for individual i , which could have a gamma or lognormal distribution. Here, the distribution of ω_i models the association between the distinct survival processes, thereby handling the correlation and enabling risk prediction of both outcomes.

To fit this model in practice, the datasets for each outcome must be stacked on top of each other, and the following shared frailty model fit to the data:

$$h(t|\omega_i) = \omega_i h_0(t) \exp(\beta_{ind} * I[X_{ind} = A] + \beta_A X_i * I[X_{ind} = A] + \beta_B X_i),$$

where $X_{ind} \in \{A, B\}$ is an indicator variable denoting which outcome the row corresponds to, β_B is the hazard ratios shared across both outcomes, and β_A tests whether the hazard ratios for outcome A differ from the hazard ratios of outcome B ($\beta_A = 0$ implies no change in hazard ratios). Note that this approach relies on the baseline hazards of the two survival processes being proportional, $h_{0,B}(t) = h_0(t)$, $h_{0,A}(t) = h_0(t) * \exp(\beta_{ind})$. To alleviate this assumption, a stratified model could be fit:

$$h(t|\omega_i) = \omega_i h_j(t) \exp(\beta_A X_i * I[X_{ind} = A] + \beta_B X_i)$$

for $j \in \{A, B\}$. This model allows the estimation of marginal risk scores given the random effect $P(T_A \leq t|X_i, \omega_i) = 1 - P(T_A > t|X_i, \omega_i)$ and $P(T_B \leq t|X_i, \omega_i) = 1 - P(T_B > t|X_i, \omega_i)$. The observations within a cluster in a shared frailty model are assumed to be independent after conditioning on the random effect, meaning the risk can be estimated as:

$$P(T_A \leq t, T_B \leq t|X_i, \omega_i) = P(T_A \leq t|X_i, \omega_i) * P(T_B \leq t|X_i, \omega_i).$$

To estimate the risk for new individuals, one needs to integrate over the distribution of ω :

$$P(T_A \leq t, T_B \leq t|X_i) = \int P(T_A \leq t|X_i, \omega) * P(T_B \leq t|X_i, \omega) f_\omega(\omega) d\omega,$$

where $f_\omega(\omega)$ is the estimated probability density function for ω .

There are some similarities, and also key differences, between frailty models and copula models which are discussed elsewhere.⁴⁶

2.6 | Multistate models

In multistate models each outcome event is seen as a state, and the probability of transitioning between different states is modeled using competing risks approaches.⁴⁰ Generally, cause-specific hazard functions are calculated for each transition, which can be done parametrically or semi-parametrically (ie, Cox proportional hazards). A diagram representation of a model for predicting two outcomes is given in Figure 1, where $h(t)$ represents the hazard rate for each transition. It is important to note that here we are working in a context of competing risks, but the goal is not to estimate competing risks scores (although this is possible within this framework, see Section 2.7). The target estimand is still $P(T_A \leq t, T_B \leq t|X)$.

After the cause-specific hazards have been fitted, risk estimates can be calculated using the transition probabilities, which is the probability of being in a given state at time t , when in a given state at time s . The risk of both outcomes can be calculated as the probability of being in the “ $A + B$ ” at time t , when in the healthy state at time 0. There are several ways to estimate the transition probabilities depending on whether the Markov assumption holds or not.⁵⁸⁻⁶² However, if all individuals start in the initial state, and one is only interested in transition probabilities out the initial state at time 0 (which is the case for this study), then the transition probabilities are equivalent to the “state occupational probabilities,” and the Aalen-Johansen (the simplest approach) will be a consistent estimator even for non-Markov data. A detailed description of how to fit multistate models using the package `mstate` is given by de Wreede et al.^{63,64}

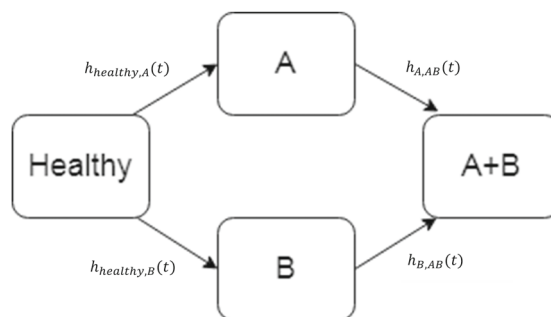


FIGURE 1 A multistate model for predicting two outcomes A and B

2.7 | Development of competing risks models

A key aspect to discuss, in the context of multiple time-to-event outcomes, is competing risks.⁴¹⁻⁴⁴ Throughout this article, we assume that the two events (A and B) do not prevent each other, motivated by examples as given in the introduction. However, there may be a third event D (e.g. death) which does act as a competing risk for both A and B . While competing risks approaches are well developed for univariate analyses, this is not the case for some of the other methods described above.

For the product method, univariate competing risks models can be fitted in R using the `mstate` package.^{40,63,64} These risks can then be multiplied to estimate the risk of both outcomes, in the same way that was done for the non-competing risk scores. Multistate models are implemented within a competing risks framework and therefore it is straightforward to produce estimates which account for a competing event by introducing an absorbing state which individuals may move into after the occurrence of the competing event. Theory on frailty models with competing risks is well developed,⁶⁵⁻⁶⁹ and more recently copula models with competing risks,⁷⁰⁻⁷⁴ but there is a lack of flexible publicly available software to implement these methods. The majority of the theory focuses on frailty terms shared between the competing risks, whereas in this study we are interested in a frailty term to estimate the risk of two outcomes both occurring in the presence of a third competing risk. We see no reason these methods would not extend to the setting in this study which is relatively simple in comparison. A straightforward method to account for a competing risk is to not censor individuals after the competing event occurs, instead setting the event time to the largest possible follow up time. They will then remain in the “at risk” group of individuals after the competing event, giving estimates of risk that account for the competing risk. The drawback of this approach is that the risk of the competing event itself cannot be calculated, which may or may not be of importance.

3 | SIMULATION

We detail the methods for the simulation using the “aims, data-generating mechanisms, estimands, methods, performance measures” (ADEMP) structure.⁷⁵ Code for running the simulation is available from our GitHub public repository.⁷⁶

3.1 | Simulation aims and overview

The first aim of the simulation was to measure the extent of the miscalibration in the risk prediction of both-of-two survival outcomes using the product method, when there was residual correlation in the outcomes. The second aim was to compare the performance (calibration and discrimination) of available methods for predicting the risk of both-of-two survival outcomes.

When interest is in predicting the risk of both-of-two outcomes, the dual-outcome approach seems the most natural approach, and motivation is required as to why the copula, frailty or msm approaches may outperform the dual-outcome. All these methods make different assumptions about the underlying data. For example, the dual-outcome relies on the survival outcome T_{AB} meeting the distributional assumptions of the chosen model. If cox regression is chosen, the hazard function for AB must meet the proportional hazards assumption, or for an accelerated failure time model, that the covariates act multiplicatively on the mean survival time. Given the complex form of the survival distribution and hazard function for the dual outcome, it is unlikely either of these assumptions would hold in practice (supporting information file 1). In contrast, the multistate model only makes these assumptions on the cause-specific hazards for each outcome in isolation. Similarly, the frailty models and copula models specify the marginal distributions separately, and then place a parametric distribution on the residual correlation. These model specifications may therefore be more appropriate than those of the dual outcome. We therefore constructed our simulation scenarios around robustness to model misspecification, to assess performance in scenarios where data was generated under a different mechanism from the model being applied. To enable a comprehensive comparison, we generated data under a variety of data-generating mechanisms, each matching the model structure of one of the analysis methods (full factorial design).

A second reason that the dual-outcome method may be outperformed by the other methods is due to statistical power. The dual-outcome approach discards some outcome event data, by ignoring events that occur on their own (ie, those patients that experience only one of the outcomes are not counted as “events”). The copula and frailty approaches estimate

predictor coefficients for each marginal distribution separately and will therefore benefit from an increase in power when estimating these predictor coefficients. This may lead to lower levels of overfitting. The multistate model may suffer from a similar issue with respect to power, as only a small number of events will occur for transitions $A \rightarrow AB$ and $B \rightarrow AB$. This may result in overfitting in the estimation of these cause-specific hazards. This is a particular risk at small sample sizes or for rarer outcomes. To assess this, we varied both the sample size and marginal risk of each outcome in the simulation.

3.2 | Data generation mechanisms

For each data generation mechanism (DGM), we simulated 1000 development datasets of size n , where n was 1000, 2500, or 5000, depending on simulation scenario. Baseline predictors were two random variables, $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$.

3.2.1 | DGM-1: Multistate model

We simulated data from the multistate model depicted in Figure 1, with exponential hazards:

$$\begin{aligned} h_{\text{healthy},A}(t) &= \left(\frac{1}{\lambda_A} \right) * \exp(X \cdot \beta_A), \\ h_{\text{healthy},B}(t) &= \left(\frac{1}{\lambda_B} \right) * \exp(X \cdot \beta_B), \\ h_{A,AB}(t) &= \left(\frac{1}{\lambda_B * \lambda_{A,B}} \right) * \exp(X \cdot \beta_B), \\ h_{B,AB}(t) &= \left(\frac{1}{\lambda_A * \lambda_{B,A}} \right) * \exp(X \cdot \beta_A). \end{aligned}$$

We assume the shape of transition $h_{\text{healthy},A}(t)$ and $h_{B,AB}(t)$ are the same as they are both transitions resulting in the development of outcome A . The term $\lambda_{B,A}$ causes a change in the scale of developing condition A (ie, the event to happen at a quicker or slower rate), once condition B has been developed. Similarly, $\lambda_{A,B}$ causes a change in the scale of developing condition B once condition A has been developed. These terms therefore controlled the level of residual correlation ($\lambda_{B,A} = \lambda_{A,B} = 1$ being no residual correlation) and are therefore key parameters to vary in this simulation. β_i is a vector containing the log-hazard ratios of the effect of X_1 and X_2 on outcome i .

3.2.2 | DGM-2 (Clayton), DGM-3 (Gumbel) and DGM-4 (Frank): Copulas

Let $C_\theta(u, v)$ be an copula. We first generated n sets of $u \sim Unif(0, 1)$, $v \sim Unif(0, 1)$ from this copula. We then generated event times from each marginal distribution separately, assuming that the random draws of u and v were from the cumulative distribution function for outcome A and B , respectively. We assumed survival times for A and B follow an exponential distribution with scales λ_A and λ_B , respectively. Specifically, the hazard for A was:

$$h_A(t) = \left(\frac{1}{\lambda_A} \right) * \exp(X \cdot \beta_A)$$

and the survival function was therefore:

$$\begin{aligned} S_A(t) &= \exp(-H_A(t)) \\ &= \exp\left(-\left(\frac{t}{\lambda_A}\right) * \exp(X \cdot \beta_A)\right). \end{aligned}$$

Then for a random event time t_A , we define $u = P(T_A > t_A)$, and calculate the simulated event time t_A as follows:

$$u = S_A(t_A),$$

$$u = \exp\left(-\left(\frac{t_A}{\lambda_A}\right) * \exp(X \cdot \beta_A)\right),$$

$$\lambda_A * \left(\frac{-\log(u)}{\exp(X \cdot \beta_A)}\right) = t_A.$$

Therefore we simulated event times t_A according to the distribution:

$$t_A \sim \lambda_A * \left(\frac{-\log(u)}{\exp(X \cdot \beta_A)}\right).$$

Event times t_B were simulated similarly.

$$t_B \sim \lambda_B * \left(\frac{-\log(v)}{\exp(X \cdot \beta_B)}\right).$$

These event times will have marginal exponential distributions with λ_A and λ_B , but will have a joint distribution that has the properties of the copula C_θ .

The term θ induces residual correlation between the outcomes, and therefore the value of θ is of key interest in simulations under this DGM.

3.2.3 | DGM-5 (log-normal) and DGM-6 (gamma): Frailty models

We first define two exponential baseline hazards for outcomes A and B , $h_{0,A}(t)$ and $h_{0,B}(t)$. For DGM-5 generate a shared Gaussian frailty term for each individual, $\omega_i \sim \text{lognormal}(1, \sigma^2)$. For DGM-6 generate a shared gamma frailty term for each individual, $\omega_i \sim \text{Gamma}\left(\frac{1}{\beta}, \beta\right)$.

We then generate survival times according to the hazard functions $h_A(t|\omega_i)$ and $h_B(t|\omega_i)$:

$$h_A(t|X, \omega_i) = \omega_i h_{0,A}(t) \exp(\beta_A X_i),$$

$$h_B(t|X, \omega_i) = \omega_i h_{0,B}(t) \exp(\beta_B X_i).$$

The term ω_i induces residual variation between the outcomes, and therefore the variance of this parameter is of key interest in simulations under this DGM. A variance of zero ($\theta = 0$ or $\beta \rightarrow 0$) will result in no residual correlation, and as the variance increases this will result in more residual correlation.

3.2.4 | Simulation scenarios and choice of input parameters

The choice of simulation scenarios was based around the aims of the simulation outlined in Section 3.1. The three major aspects of the simulation that were varied were (i) DGM; to assess each methods sensitivity to model misspecification, (ii) the level of residual correlation; this quantity drives the bias in the product method and is the reason we must use these alternative approaches, and (iii) the amount of statistical power available to predict the outcome; this may also drive model performance.

We have outlined the 6 DGMs in Sections 3.2.1 to 3.2.4. We created a further six scenarios: lower marginal risks, no residual correlation (LN); lower marginal risks, lower residual correlation (LL); lower marginal risks, higher residual correlation (LH); higher marginal risks, no residual correlation (HN); higher marginal risks, lower residual correlation (HL); and higher marginal risks, higher residual correlation (HH), based off the amount of residual correlation and incidence of the outcomes.

TABLE 1 Targeted values of marginal risks and risks of both outcomes occurring for each simulation scenario

	Scenario LN	Scenario LL	Scenario LH	Scenario HN	Scenario HL	Scenario HH
Marginal risk <i>A</i>	10%	10%	10%	30%	30%	30%
Marginal risk <i>B</i>	10%	10%	10%	30%	30%	30%
$j_{r_{ind}}$	1%	1%	1%	9%	9%	9%
$j_{r_{pred}}$	1.25%	1.25%	1.25%	11.25%	11.25%	11.25%
$j_{r_{true}}$	1.25%	1.5%	1.875%	11.25%	13.5%	16.875%
$j_{r_{true}}/j_{r_{pred}}$	1 (0% increase)	1.2 (20% increase)	1.5 (50% increase)	1 (0% increase)	1.2 (20% increase)	1.5 (20% increase)

Note: $j_{r_{ind}}$ = mean risk in population assuming complete independence (no conditioning on predictors); $j_{r_{pred}}$ = mean risk in population assuming independence after conditioning on available predictors; $j_{r_{true}}$ = true mean risk in population; process for calculating $j_{r_{ind}}$, $j_{r_{pred}}$, and $j_{r_{true}}$ are given in supporting information file 1.

The lower marginal risk scenarios (LN, LL, and LH) were targeted to have marginal risks of outcome *A* (10%) and *B* (10%) in line with the marginal risks of the outcomes in the clinical example from Section 4. These were increased to a marginal risk of 30% for each outcome in the higher marginal risk scenarios (HN, HL, and HH). The lower residual correlation scenarios (LL and HL) were targeted to have a true mean risk 20% higher than the mean risk when assuming independence after conditioning on predictor variables. This was increased to 50% for the higher residual correlation scenarios (LH and HH). The targeted values for each scenario are displayed in Table 1. We aimed to keep the magnitude of predictor coefficients similar across each DGM within a given scenario. The exact input parameters for each scenario, and a more detailed explanation of the process for choosing these parameters, are provided in supporting information file 1.

Finally, we created more scenarios based on development dataset sizes of $n \in \{1000, 2500, 5000\}$. The minimum sample size (1000) was estimated using the sample size formula of Riley et al,⁷⁷ for a risk prediction model predicting the dual-outcome in scenario LN. A conservative estimate of $R^2_{CS_{adj}}$ equivalent to an $R^2_{Nagelkerke}$ of 0.15 was used in the calculation, giving a minimum required sample size of 896. We therefore chose $N = 1000$ as the smallest sample size in the simulation. Code for this step is provided on GitHub.⁷⁶ This resulted in a total of $6*4*3 + 2*3 = 78$ scenarios (note that for no residual correlation scenarios LN and HN we do not generate data using every DGM). The input parameters we have chosen to vary are based on the aims of the simulation. A censoring time was simulated from a survival distribution with an exponential hazard, and log hazard ratios of 0.1 for both X_1 and X_2 . The rate was chosen to target 5% of the events to be censored in the lower marginal risk scenarios (see supporting information file 1 for exact values).

When interpreting the results, it is important to note that for a given DGM, each misspecified model may be a different “distance” away from the model used to generate the data. If two models are very dissimilar, we would expect poor performance for either of these models when the other is used for the DGM. If a given model has very poor performance across a range of the DGMs, it is likely quite “far” from all the other model structures, and will be deemed sensitive to model misspecification.

3.3 | Estimands and other targets

The main estimand was the set of points $P(T_A \leq 3653, T_B \leq 3653|X)$ over all individuals in the validation cohort. This is the risk of developing outcome *A* and *B*, prior to time $t = 3653$, which corresponds to 10-year risk (assuming an integer to be 1 day) in line with the clinical example. A further target of the simulation was to report on the discrimination of each model in the validation cohort.

3.4 | Methods

Each of the methods outlined in Section 2 was used to estimate $P(T_A \leq t, T_B \leq t|X)$, as described in Section 2. The estimated risks of developing both *A* and *B* are hereto referred to as $risk_{product}$ (product of univariate models), $risk_{d-o}$

(dual-outcome model), $risk_{cop-clay}$, $risk_{cop-gum}$, $risk_{cop-FGM}$ (copula model assuming Clayton, Gumbel, or FGM copulas), $risk_{frail-norm}$, $risk_{frail-gam}$ (frailty model assuming log-normal or gamma frailty distribution), $risk_{msm}$ (multistate model).

The dual-outcome model was fitted using a Cox proportional hazards model given this is the most common approach in practice. It was therefore important to test its performance. Alternative parametric approaches could be used to remove reliance on the proportional hazards assumption, but will each have their own set of distributional assumptions. The frailty models were fitted using a Bayesian MCMC approach utilizing the rstan package.⁷⁸ A Weibull baseline hazard was assumed, but any distribution could be used. Code for this is available from our GitHub public repository.⁷⁶ This approach was used as the likelihood was very flat and convergence issues were encountered when attempting to fit these models using maximum likelihood or expectation maximization algorithms.

3.5 | Performance measures

We compared each method's ability to estimate the set of points $P(T_A \leq 3653, T_B \leq 3653|X)$ by assessing moderate calibration.⁷⁹ Calibration was assessed in a validation cohort of size 1000 generated using the same DGM as the development dataset. A new validation dataset was simulated for every development dataset. We generated flexible calibration curves by regressing the true risks on the predicted risks using a linear model with multiple fractional polynomials. True risks under each DGM were calculated using the process given in supporting information file 1. While in practice one would regress on the observed outcomes themselves, the simulation allows us to regress on the underlying true risk, thereby giving a more accurate assessment of calibration. The resulting curves gave the observed (true) risk as a function of predicted risk. We report the (pointwise) median and 5th/95th percentile of the calibration plots across the 1000 simulation iterations. Throughout this article, we refer to the median calibration curve and 5 to 95 percentile range in calibration curves across the 1000 iterations as "average calibration" and "calibration variation," respectively. A detailed process for producing these calibration plots is given in supporting information file 1. Discrimination was assessed in the validation cohort using Harrell's C statistic.⁸⁰

There are two points of clarification with regards to calibration as a performance measure. First, note that calibration is defined as the difference between the predicted risks and observed risks (or event rates) in a cohort of interest. In this simulation, the estimand itself are the "observed risks," and the predicted risks from each method are the "predicted risks." We therefore reason that the average calibration curve in this simulation is analogous to bias. Similarly, the calibration variation is analogous to the SE of the estimator of the risk. Second, note that the estimand in this simulation is a set of points. Therefore rather than presenting the bias in the estimation of a single estimand, our bias is presented as a line over the range of predicted risks. There is also a different level of variability at each point along this line. For example when individuals have very low predicted risks (near 0), we will expect to see less calibration variation than when predicted risks are bigger.

3.6 | Simulation results

3.6.1 | Discrimination

For all scenarios the discrimination of all methods were similar (supporting information file 2: Tables S4.1-S4.14), meaning each methods ability to risk-rank individuals was similar. For scenarios LN and LL, $N = 1000$ only, there was a small drop in the discrimination of the dual-outcome and msm methods. This simplifies the discussion greatly. Both calibration and discrimination are seen as highly valuable performance metrics to report.^{1,2,81,82} However each may be more important in different clinical settings. For example, if the clinical strategy is to treat all individuals over a certain risk threshold with a low-cost intervention (such as statins), then calibration may be more important than discrimination. However, if there is limited resources and an intervention can only be given to a fixed number of individuals who would all benefit from the treatment (say a diagnostic operation), then discrimination is arguably more important, to ensure the highest risk individuals receive the treatment first. Given the similar level of discrimination of every method, we do not have to weigh up the importance of calibration vs discrimination, as there are no scenarios where a method outperforms the others with respect to calibration but performs worse with respect to discrimination. We therefore focus on the calibration for the remainder of the results section.

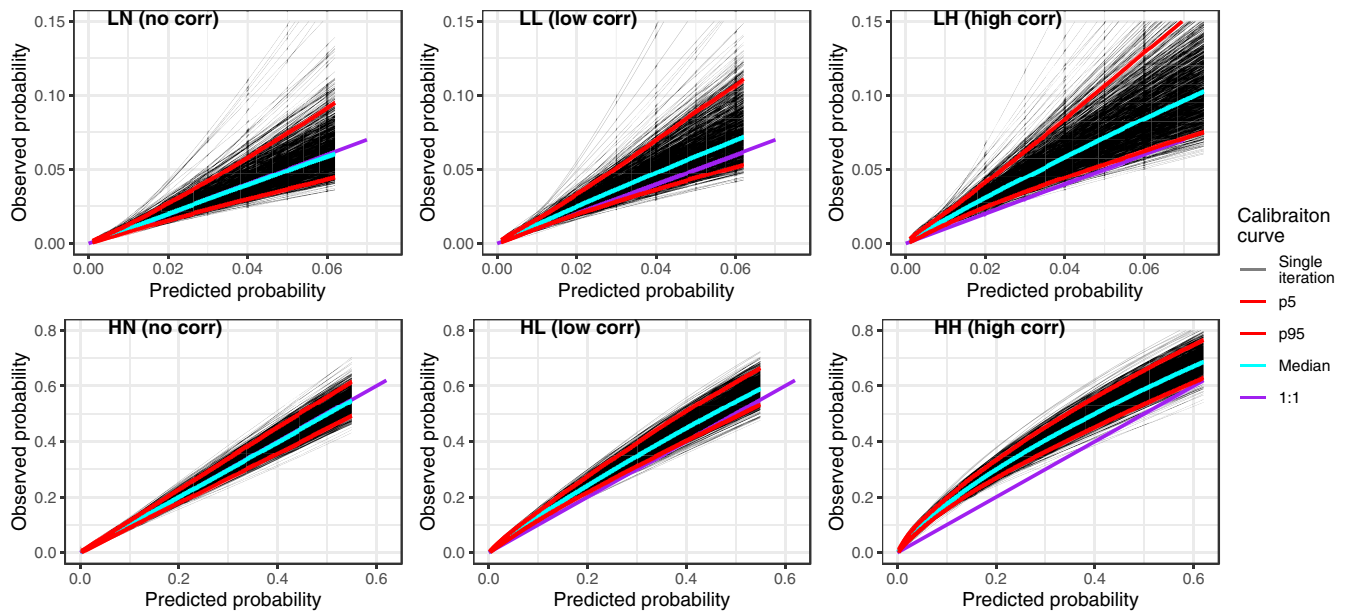


FIGURE 2 Calibration curves of the product method across the 1000 simulations, paneled by simulation scenario ($N = 1000$). Blue line = median calibration curve; red lines = fifth and 95th percentile in calibration curves. For scenarios LL, LH, HL and HH, the Clayton DGM (DGM-2) has been used

3.6.2 | Calibration

To answer the first aim of this study, we present the calibration curves of the product method for varying levels of residual correlation (Figure 2). When there was no residual correlation (scenarios LN and HN), the product method was well calibrated on average across the entire range of predicted risk. However as residual correlation increased (scenarios LL, LH, HL, and HH) the product method became increasingly miscalibrated, often underpredicting the risk. This was consistent regardless of the sample size of the development cohort (supporting information file 2: Figures S1.1-S1.3).

To answer the second aim of this study, we now present the average calibration, and calibration variation of all the analysis methods over a range of scenarios and sample sizes. We have selected a subset of results to present in the main article, however equivalent plots for all scenarios and sample sizes are available in supporting information file 2. We present the results in line with key aspects of the simulation that were outlined in Section 3.2.4. For the rest of this section, when we refer to “all methods,” we are not considering the product method.

Impact of model misspecification

The dual-outcome and msm methods had the most consistent performance across all the DGMs, indicating they were the most robust to model misspecification. For example, in scenario LL when $N = 5000$ (Figure 3), the msm and dual-outcome methods had extremely good average calibration across every DGM, demonstrating their ability to handle model misspecification. The Clayton and Frank copula models also had very good average calibration for 5 of the DGMs but were very poorly calibrated under the Gumbel DGM. On the contrary, the Gumbel copula method was very poorly calibrated under all DGMs except the Gumbel DGM. The normal and gamma frailty models had good average calibration across all DGMs, but not as strong as the msm or dual-outcome methods.

For scenario LL when $N = 1000$ (Figure 4), the average calibration of the msm and dual-outcome methods were poor across all DGMs (more so for the msm) and showed signs of overfitting; under prediction at lower predicted risks and over prediction at higher predicted risks. However, it should be noted that performance was consistent across all DGMs, supporting the idea that the both models are robust to model misspecification. The average calibration of the other methods (Clayton, Frank and Gumbel copula models, and normal and gamma frailty models) had a similar pattern to when $N = 5000$. All methods (except the Gumbel copula) had poor average calibration under the Gumbel DGM and had good

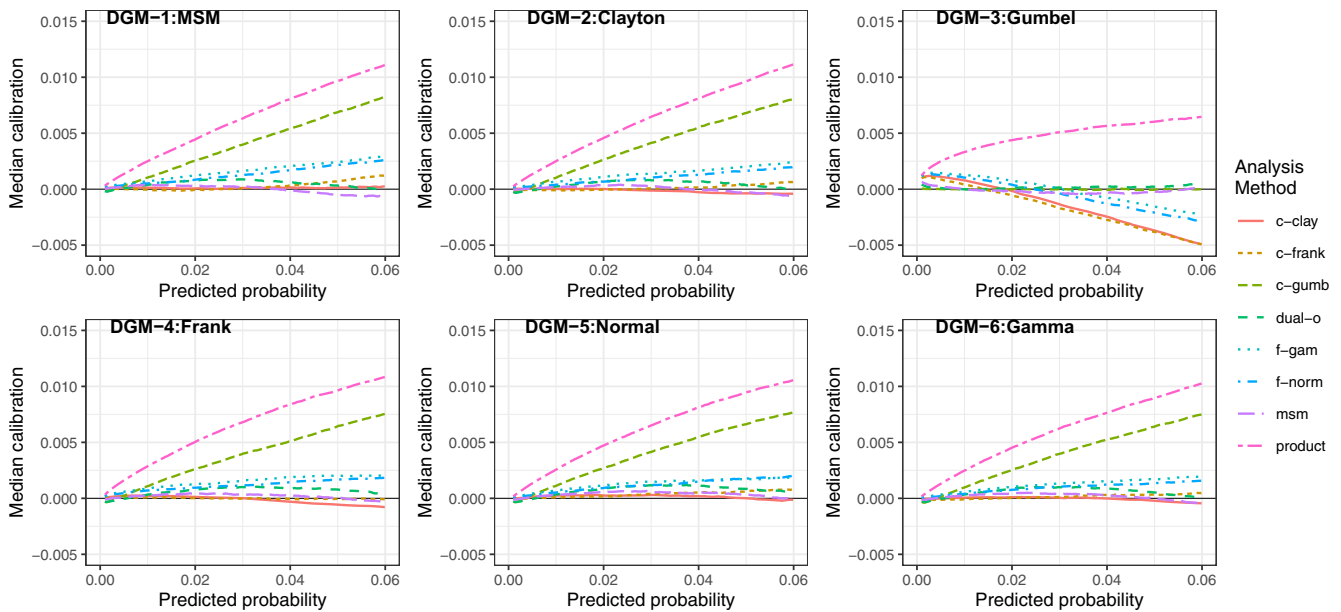


FIGURE 3 Median calibration (observed–predicted risk) curves across the 1000 simulations for scenario LL (lower outcome prevalence, lower residual correlation), $N = 5000$

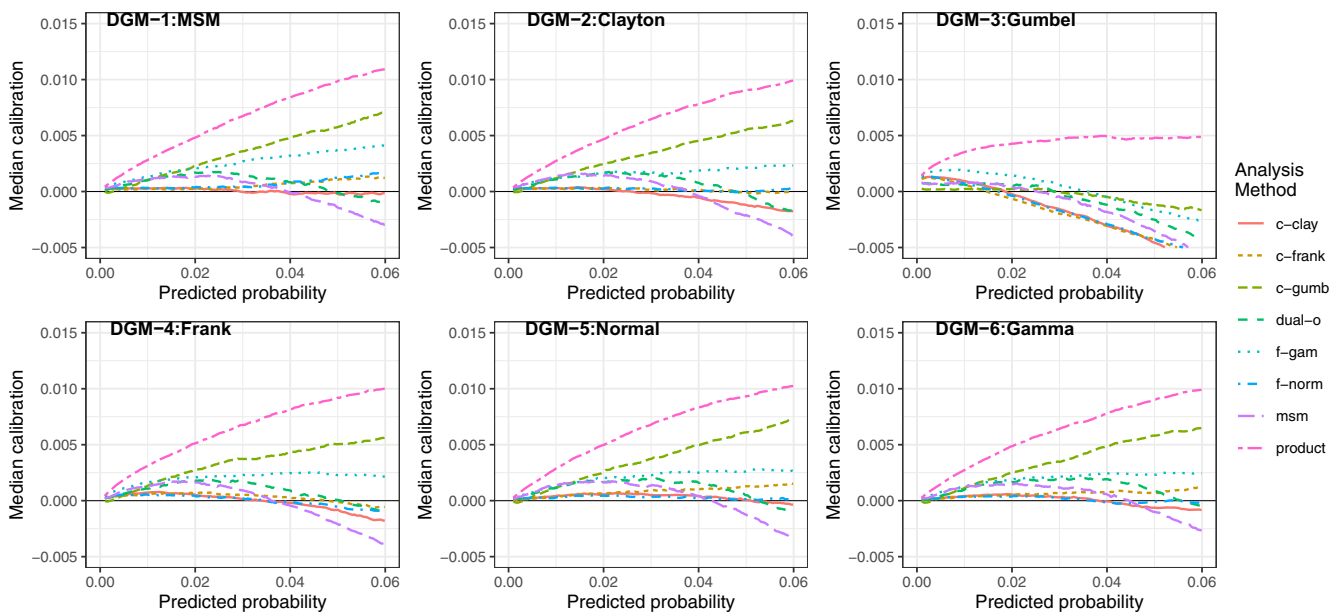


FIGURE 4 Median calibration (observed–predicted risk) curves across the 1000 simulations for scenario LL (lower outcome prevalence, lower residual correlation), $N = 1000$

average calibration across the other 5 DGMs, with performance particularly strong when each model was correctly specified. The Gumbel model had very poor average calibration under all DGMs except the Gumbel DGM. This indicates that the frailty and copula methods are more sensitive to model misspecification, in particular the Gumbel copula.

Impact of increasing sample size or marginal risks

As noted in the previous section, as the sample size decreased (Figures 3 and 4) the msm and dual-outcome were more prone to overfitting than the frailty or copula models. The similarity in average calibration between $N = 1000$ and 5000

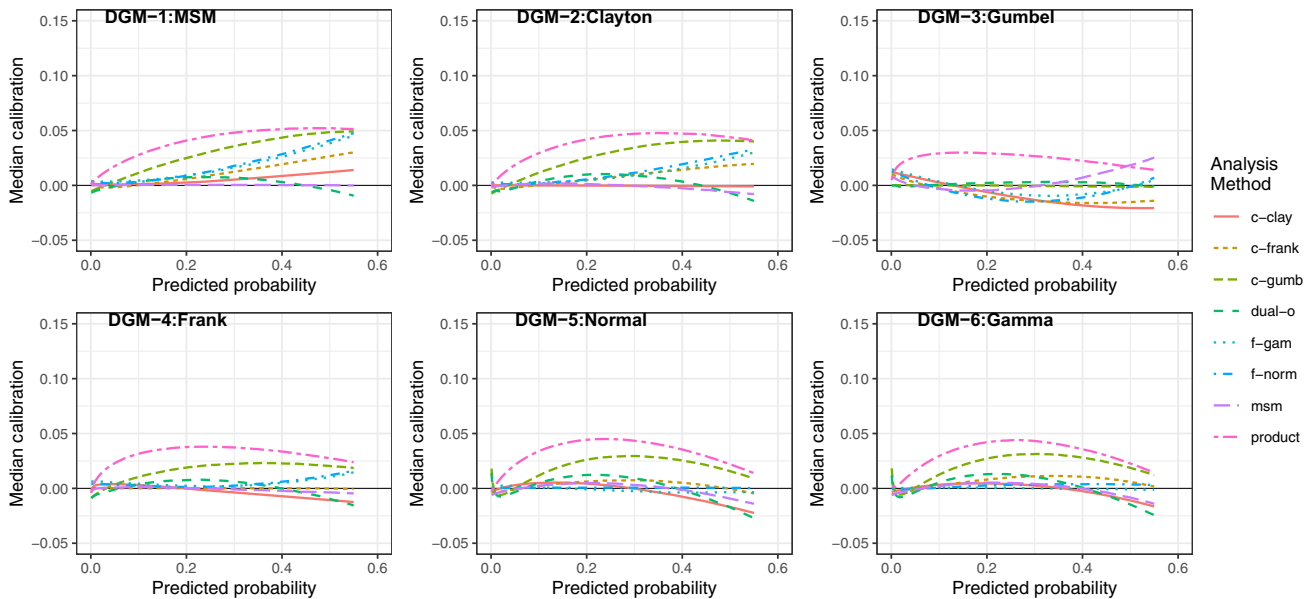


FIGURE 5 Median calibration (observed–predicted risk) curves across the 1000 simulations for scenario HL (higher outcome prevalence, lower residual correlation), $N = 5000$

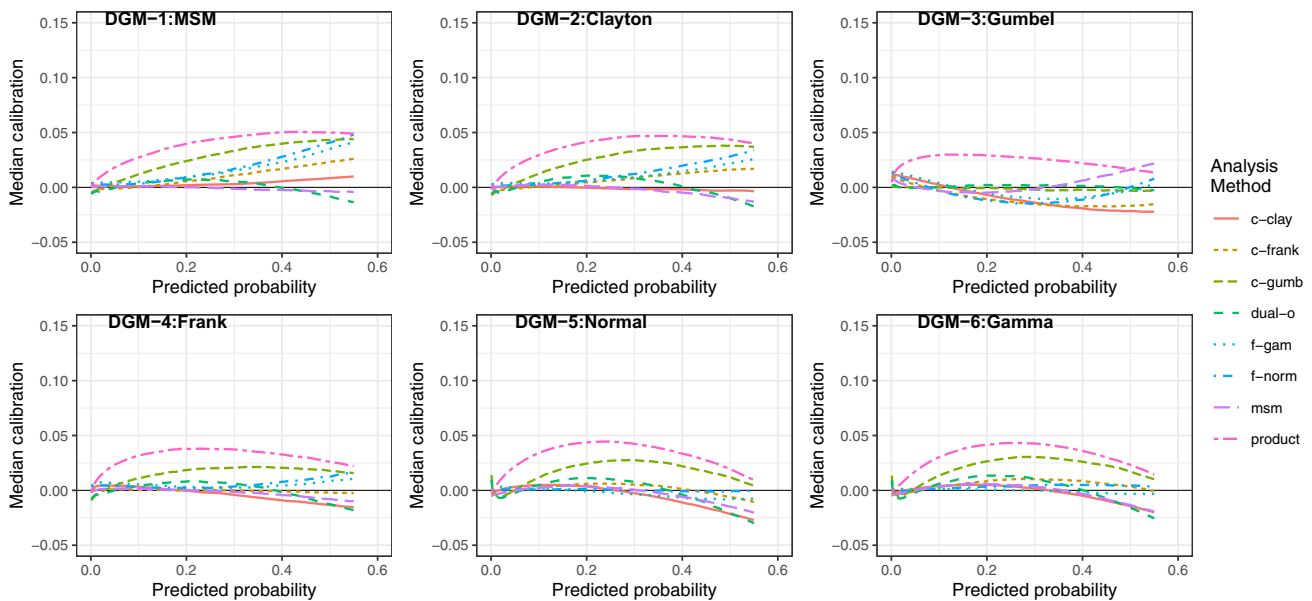


FIGURE 6 Median calibration (observed–predicted risk) curves across the 1000 simulations for scenario HL (higher outcome prevalence, lower residual correlation), $N = 1000$

for the frailty and copula methods indicates that while they are more sensitive to model misspecification, they are less data intensive approaches.

When the incidence of both outcomes increased (but keeping the same level of residual correlation)—that is comparing scenario HL (Figures 5 and 6) with scenario LL (Figures 3 and 4)—all methods were generally better calibrated at lower sample sizes. This was highlighted by there being less of a difference in performance between $N = 5000$ (Figure 5) and $N = 1000$ (Figure 6); that is, the required sample size to mitigate overfitting decreased. When the marginal incidence was higher, the msm and dual-outcome were less impacted by the smaller sample size (Figure 6), retaining good calibration for $N = 1000$.

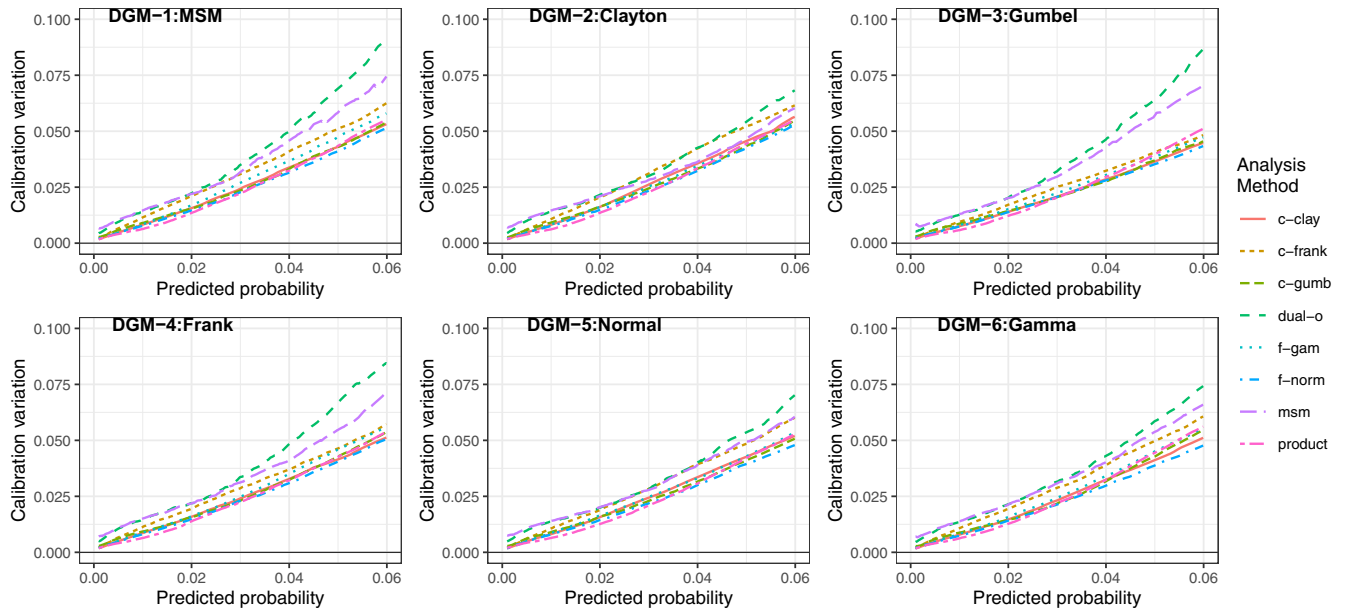


FIGURE 7 5 to 95 percentile range of calibration curves across the 1000 simulations for scenario LL (lower outcome prevalence, lower residual correlation), $N = 1000$

Figure 7 plots the 5 to 95 percentile range in calibration curves of each method against the predicted risk in scenario LL at $N = 1000$. The value of the y-axis is analogous to the distance between the blue lines (fifth percentile and 95th percentile) in Figure 2. As previously mentioned, the level of variation in the calibration curves increases when the predicted risks get bigger. The msm and dual-outcome had the largest variability in calibration across every DGM, with little to choose between the other methods. The Frank copula method generally had the third highest calibration variation. This greater calibration variation for msm and dual-outcome supports the idea that they are more data intensive approaches. These results held across all scenarios and sample sizes (supporting information file 2: Figures S3.1-S3.12), although as the sample size increased the absolute difference in calibration variation between all analysis methods became negligible.

Impact of increasing the residual correlation

We saw similar patterns in results for the scenarios with higher levels of residual correlation (supporting information file 2: Figures S2.4-S2.6 and S2.10-S2.12). The dual-outcome and msm methods again had the most consistent median calibration across all DGMs and performed well at the higher sample sizes. The Clayton and Frank copula methods also had good median calibration, consistent with that of dual-outcome and msm for the majority of DGMs. There were very high levels of miscalibration in the product method, to be expected given the higher levels of residual correlation.

Figure 8 contains the average calibration of each method when there was no residual correlation present. For the higher marginal incidence scenario (HN) all the methods had near perfect average calibration (slight deviations for the dual-outcome method). For the lower marginal incidence scenario (LN) there was greater differences in the performance of each method. We found that the product method and Frank copula methods had the best average calibration across sample sizes, however at $N = 5000$ the dual-outcome and msm models also performed well. The Clayton and Gumbel copulas and both frailty models over predicted risk at all sample sizes.

4 | CLINICAL EXAMPLE

4.1 | Aims and setting

The aim of this clinical example was to assess the performance of each of the methods outlined in Section 2 in a real clinical setting. We considered the prediction of the 10-year risk of CVD and T2D developing together. The impact of multimorbidity on healthcare systems and patient outcomes has been well documented,²⁷⁻³⁴ as well as an increased economic

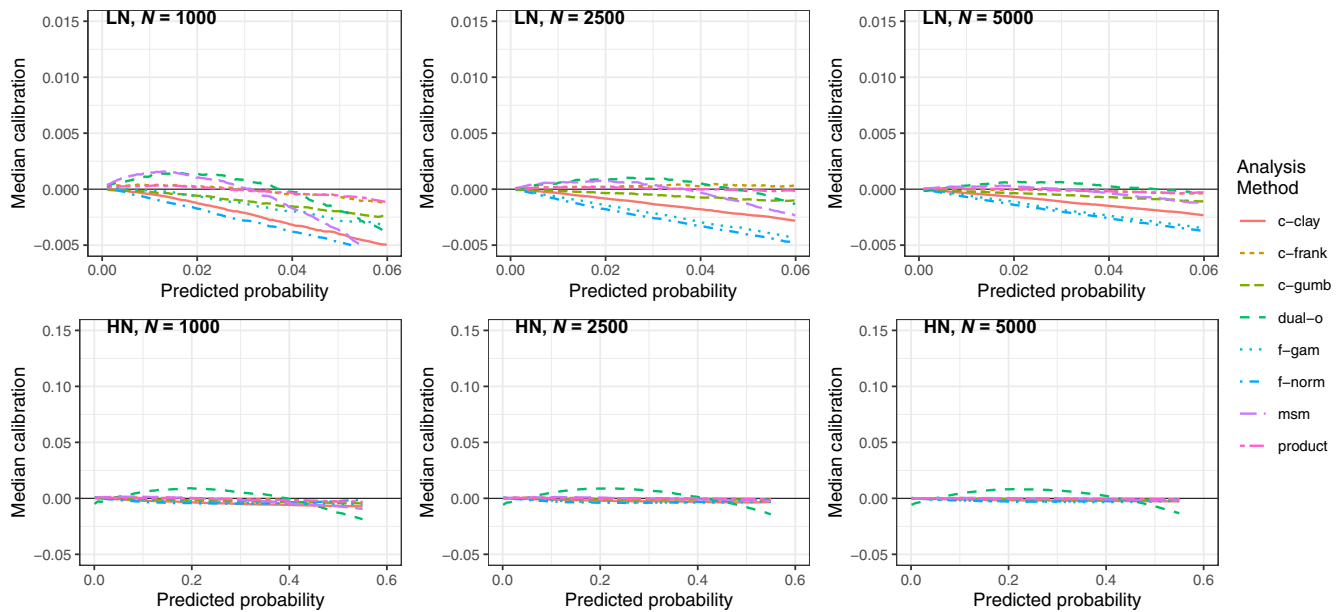


FIGURE 8 Median calibration curves (observed–predicted risk) across the 1000 simulations in the presence of no residual correlation (scenarios LN and HN), $N = 1000, 2500, 5000$. All the DGMs were equivalent in the presence of no residual confounding and so plots are not separated by DGM

burden specifically for those with CVD and T2D,^{83,84} and increased levels of mortality.^{85,86} Knowing the risk of developing both of these conditions for specific groups of individuals in the population would enable health care providers to optimize resource allocation.

4.2 | Methods

4.2.1 | Data source

Data from the Clinical Practice Research Datalink (CPRD), linked to admitted patient care data from Hospital Episode Statistics (HES) and death data from the Office for National Statistics (ONS), was used to build these models. CPRD GOLD and CPRD Aurum are primary care datasets containing data from general practices with the Vision and EMIS Web computer systems, respectively.⁸⁷ CPRD Aurum was used in this study, which covers practices in England and Northern Ireland, with >39 million historical patients, and >13 million currently registered. It is representative of the English population in terms of age, gender, geographical spread and deprivation (as of 2019).⁸⁸

4.2.2 | Outcomes and predictors

We extracted a cohort including all patients that had at least 1 day of follow up in the database aged >65 after 1 January 2000, and at least 1 year up to standard registration prior to this point. Start of follow up was defined as the maximum of date turned age 65, 1 January 2000, and date of 1 year of up to standard registration in the database. End of follow up was defined as the minimum of date of death, transferred out of practice, or last data collection for practice. Individuals were then excluded if they had a history of CVD or T2D event prior to their start of follow up. CVD and T2D events were identified through the CPRD, HES, and ONS data sources. CVD was defined as a composite event consisting of heart failure, myocardial infarction, coronary heart disease, stroke, and transient ischaemic attack. We considered the following predictors at start of follow up for each individual: age, gender, smoking status (never, ex-smoker, current smoker), systolic blood pressure (SBP), cholesterol/high density lipoprotein (chol/HDL) ratio, index of multiple deprivation (IMD), body

mass index (BMI), and ethnicity (Black, Chinese and other, Mixed race, South Asian, White). We included all variables included in the SCORE risk prediction model (used for CVD risk assessment across Europe), plus any test data variables used as predictors in QRISK3 (used for CVD risk assessment in England and Wales). Predictor variables were identified through CPRD only. Code lists for all variables and algorithms for extracting test data are provided on GitHub.⁷⁶ Operational definitions for extracting all variables and details on the code-lists are given in supporting information file 1.

4.2.3 | Data preparation

$N = 2\,074\,323$ individuals met the inclusion/exclusion criteria and were included in the cohort. There were missing data on Smoking status, SBP, chol/HDL ratio, BMI, IMD, and ethnicity. We wanted to focus on what happened without missing data, we therefore created a pseudo “complete” case dataset by imputing missing values using a single stochastic imputation, obtained through a single multiple imputation chain.⁸⁹ In practice we recommend implementing a full multiple imputation. $N_{dev} = 100\,000$ and $N_{val} = 100\,000$ individuals were then selected at random for the development and validation cohorts. More details on the imputation process, including convergence and density plots for all imputed variables, are provided in supporting information file 1.

4.2.4 | Data analysis and performance measures

Let $T_{CVD,T2D}$ be the time until both CVD and T2D have occurred. Models were developed in the development cohort to predict the 10-year risk, $P_{CVD,T2D} = P(T_{CVD,T2D} < 3652.25|X)$, using each of the methods outlined in Section 2. This was done using the survival package⁹⁰ (product method and dual-outcome method), GJRM⁵² (copula models), rstan⁹¹ (frailty models) and mstate^{63,64} (multistate model). The frailty models were fit using Bayesian statistical inference and Monte Carlo Markov chains, assuming a Weibull marginal baseline hazard for each outcome. Prior distributions, the distributions from which initial values were drawn from and convergence plots are provided in supporting information file 1. For the Clayton and Gumbel copula models, we tested rotations of the copula of 90°, 180°, and 270° (rotations were not possible with the Frank copula). Calibration of each was assessed visually and the best fitting copula was used in the final analysis.

Predicted risks $\hat{P}_{CVD,T2D}$ were then generated for each individual in the validation cohort. We assessed calibration using graphical calibration curves.⁹² To do this, the complementary loglog transformation of $\hat{P}_{CVD,T2D}$, $\widehat{CLOGP}_{CVD,T2D} = \log(-\log(1 - P_{CVD,T2D}))$, was used as the sole predictor in a cox proportional hazards model, predicting $T_{CVD,T2D}$:

$$h_{CVD,T2D}(t) = h_0(t) * \left(rcs \left(\widehat{CLOGP}_{CVD,T2D} \right) \right),$$

where $h_{CVD,T2D}(t)$ is the hazard function for $T_{CVD,T2D}$, and rcs denotes restricted cubic splines (5 knots) on the predictor variable. Observed risks were then estimated by estimating a baseline hazard function for this model and calculating fitted values for each individual in the validation cohort using this model.

This approach places an assumption of proportional hazards on the outcome with respect to the complementary log-log transformation of the predicted risks, which may not be valid. We allow some deviation from this assumption by introducing cubic splines. We therefore also split the validation cohort into deciles of predicted risk and calculated the average predicted risk, and a Kaplan-Meier estimate of observed risk within each decile, to give a binned calibration plot.⁹³ While this approach has its own limitations (categorization of a continuous variable resulting in loss of information), this is a nonparametric way of assessing calibration and provides an alternative assessment of calibration. Discrimination was assessed using Harrell’s C .⁸⁰

4.3 | Results

Baseline data on development and validation cohorts are provided in supporting information file 2: Table S6.1. Figure 9 contains graphical calibration curves for each method. We have plotted over the majority of the density of the predicted risk, but not the full range, to allow a more granular comparison of the methods. See supporting information

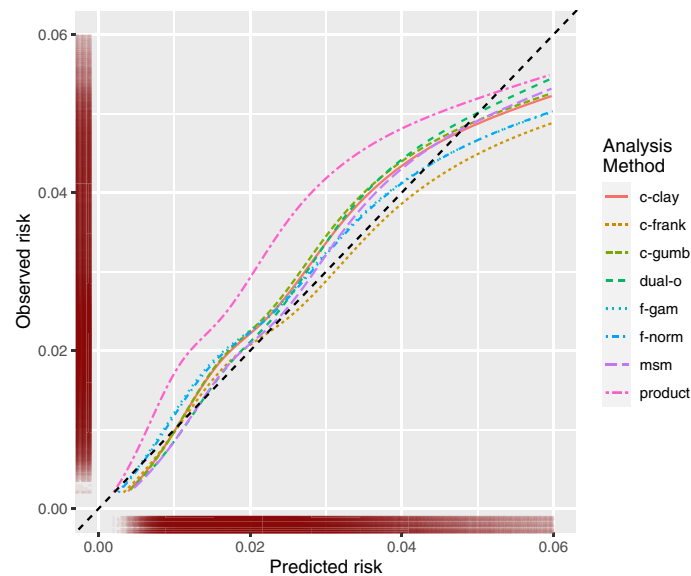


FIGURE 9 Graphical calibration curves of each method in the clinical example

file 2: Figure S6.1 for a plot over the full range of predicted risk. The product method was the worst calibrated, often underpredicting risk. In comparison, all the other methods were well calibrated, although suffered from over prediction at higher predicted risks. Of the remaining methods, the Frank copula had the best calibration up to a predicted risk of 0.04 but suffered the most from the over prediction of risk at the higher risk values. The msm had the next best calibration. The dual-outcome method had poor calibration between predicted risk of 0.03 to 0.04, but had very good calibration below 0.03, and suffered the least from over prediction at the higher end of predicted risks. Moderate calibration assessed by observed (Kaplan-Meier) vs predicted risk within deciles of predicted risk is presented in supporting information file 2 (Figure S6.2). This again showcases that the dual-outcome method was the least affected by the extreme values, with the highest risk decile being perfectly calibrated. In these plots, dual-outcome and msm were also the only methods where the observed risks increased for each decile of each predicted risk. This highlights a level of miscalibration not picked up by the graphical calibration curves. Harrel's C-statistic of all methods was the same (0.69).

5 | DISCUSSION

5.1 | Summary from the simulation

The product method had very poor calibration in the presence of residual correlation and had the worst average calibration of all the analysis methods across all scenarios. We therefore do not recommend using this approach in practice. The msm and dual-outcome methods were the most robust to model misspecification as they were the only methods to have similar levels of performance across all the DGMs. This is likely because they do not make parametric assumptions on the distribution of the residual correlation, unlike all the other methods. For larger sample sizes they were also the best calibrated methods, meaning if the sample size is sufficient, these would be the most appropriate methods to use in practice. On the contrary, these methods were the most prone to miscalibration issues at small sample sizes. They also had the highest levels of variability in calibration across simulation iterations (although the differences between each method became negligible at bigger sample sizes). As the marginal incidence of each outcome increased the performance of all the methods became more similar (although the product method and Gumbel copula still had the worst calibration across all the scenarios). The dual-outcome and msm approaches performed well even at small sample sizes in this context, the higher marginal incidences overcoming their power limitations.

5.2 | Summary from the clinical example

Considering the very poor calibration of Frank copula at the higher range of predicted risk, dual-outcome and msm had the best calibration. Alike to the simulation, the discrimination of all methods was the same. From a computational perspective (which was only an issue due to the large sample size of 100 000 used in this clinical example) the dual-outcome and product method models were fit quickly, whereas the copula models took a couple of hours, and the frailty models both took approximately a week to run, with each MCMC being run in parallel. It is possible this runtime could be reduced by better choice of priors and starting values. The msm model also took over a week to run, largely driven by the time it takes to generate predicted risks for each individual in the validation cohort. The dual-outcome and product methods could therefore be easily extended to model more predictor parameters and a higher number of individuals, which is not the case for the other methods.

5.3 | Overall discussion

Aim 1, to measure the extent of the miscalibration in predicting the risk of both-of-two survival outcomes using the product method, when there is residual correlation in the outcomes: If predicting the co-occurrence of multiple time-to-event outcomes is of interest, then a technique that models the residual correlation is essential to obtain a well calibrated estimate of risk in the presence of residual confounding. This was highlighted by the poor calibration of the product method throughout the simulation in this study. Furthermore, the product method was found to be miscalibrated in a real clinical setting predicting the risk of CVD and T2D both occurring. This is an important finding because the product method is only biased if the conditional independence assumption does not hold. One may therefore argue that when adjusting for major risk factors with established biological mechanisms, that the bias of the product method may be negligible. However, we found that when adjusting for 8 major cardiovascular risk factors, which are common risk factors for T2D too, there was still a significant level of miscalibration and under prediction of the risk.

Aim 2, to compare the performance of available methods for predicting the risk of both-of-two survival outcomes: To estimate the risk of both-of-two survival outcomes in the presence of residual confounding we recommend the dual-outcome approach. This method, along with the msm, was the most robust to model misspecification, had good calibration across a wide range of scenarios and performed well in the clinical example. This method is also the most practical as it can be implemented using standard survival analysis techniques. This means nonlinear modeling of predictors,⁹⁴⁻⁹⁹ accounting for competing risks⁴⁰⁻⁴⁴ and implementation of variable selection^{100,101} can be easily achieved using existing methodology. How to approach these topics within the framework of the other modeling approaches is currently unclear. If discrimination is the most important performance metric on which a model is being assessed, our simulation indicates all methods are equally valid options. However, we still recommend the dual-outcome approach given its robustness to model misspecification in terms of calibration.

However, there are drawbacks to this approach which must be made clear. First, this method suffered from a drop in calibration performance at small sample sizes driven by overfitting. Here, the copula (in particular Clayton and Frank) and frailty models sometimes gave better calibration, yet which model had the best calibration was dependent on the DGM. This highlights that at small sample sizes it becomes more important to understand the underlying data structure and avoid model misspecification. On the contrary, while the calibration of the dual-outcome method was poor at small sample sizes, it was consistent across all DGMs, highlighting the method's robustness to model misspecification. Poor calibration induced by overfitting can be mediated by reducing the number of predictors and ensuring sample size criteria are met,⁷⁷ or applying shrinkage and penalisation techniques. When developing a model on a small development dataset this may be an alternative approach, however would come at a cost to the discrimination of the model, which must therefore be weighed up against the risks of misspecifying the model if using a copula or frailty model.

Another drawback of the dual-outcome approach is that it does not give a direct measure of the level of association between the two outcomes. On the contrary, all of the other methods provide a quantifiable assessment of the level of dependence between the two outcomes: the estimated distribution parameter of the copula models (θ); the variance of the random effect ω_i in the frailty models; and the estimated increase in the hazard rate after conditions have been developed in the msm ($\lambda_{A,B}$ and $\lambda_{B,A}$). The dual-outcome method likely has the best performance in our study as it directly targets the estimand that we are interested in, whereas the other approaches are restricted into parametric structures designed to estimate the level of dependence, which may not be appropriate. It is therefore important to establish the primary research question and potential uses of the model when choosing the modeling approach. While the dual-outcome model is most

TABLE 2 Summary of main conclusions from the study

Simulation aspect	Conclusion
Changing DGM (evaluating robustness to model misspecification)	<ul style="list-style-type: none"> The dual-outcome and msm approaches were the most robust to model misspecification. They had consistent performance across all DGMs, whereas the frailty and copula models were more sensitive to the choice of DGM.
Increasing the level of residual correlation	<ul style="list-style-type: none"> Increasing the level of residual correlation had no impact on model performance of the methods relative to each other (excluding the product method). Dual-outcome and msm were the most robust to model misspecification for the lower and higher levels of residual correlation.
Increasing the sample size or marginal risks	<ul style="list-style-type: none"> The dual-outcome and msm were the most prone to overfitting at lower sample sizes and when marginal risks were lower. Even when impacted by overfitting, the dual-outcome and msm approaches were still the most robust to model misspecification
Other	<ul style="list-style-type: none"> Discrimination of all methods was very similar. In the clinical example, the product method resulted in a miscalibrated estimate of the risk of CVD and T2D, despite conditioning on 8 major risk factors. This suggests the conditional independence assumption may not hold in practice and motivates the use of the methods considered in this study.

suitable for estimating $P(T_A \leq t, T_B \leq t|X)$, if any interest lies in estimating the dependence structure other models must be considered at the potential cost of some predictive performance.

There are a number of other points of note to raise about each of the methods. Unlike all the other methods, the dual-outcome model does not also allow estimation of the marginal risks within a common framework. Arguably this is not an issue, as developing separate survival models to estimate these would be no more complex than implementing the other proposed approaches. The frailty approach has an added benefit that it can be fit to datasets with individuals who are missing either outcome A or B (but not both). The correlation in the observed data can be used to recover some of the missing information, which is a more common scenario where multivariate methods may be used. It is unclear whether this would be preferable over using an “impute then model” type approach, or even running a complete case analysis, and further work is needed here.

Table 2 contains a summary of the main conclusions from the study. The conclusions are categorized based on the three major aspects of the simulation that were varied (outlined in Section 3.2.4) and an “other” category.

Limitations: We focused only on two outcomes to serve as initial work in this space; however, all the methods considered could be used to predict the risk of more than two time-to-event outcomes. Further research will be needed to explore their behavior as the number of outcomes increases. We also assumed a common censoring mechanism for both outcomes, although all the methods compared in this study can be applied when the two outcomes have different censoring mechanisms, and we can hypothesize no reason why this would affect performance. Despite this, it may be worthwhile exploring this in future work. As is the case for all simulations, it is possible that the data on which our simulation was based is not representative of real clinical data on which the methods would be used in practice. We tried to alleviate this by considering a range of DGMs, 6 in total, and made each modeling approaches ability to perform across this range of DGMs a key aspect of the simulation and how we interpreted the results. We also considered a range of scenarios, of which the primary one (scenario LL), was matched to our clinical example. Furthermore, we compared the performance of each method in a real clinical example, and we were able to confirm the fallibility of the product method in this setting. The strong calibration of both the dual-outcome method and msm method in the clinical example further validates the findings from the simulation. Finally, one of the main findings was the relative impact of overfitting on the dual-outcome approach compared to the others at small sample sizes. However, we did not implement any shrinkage or penalisation techniques, which may alleviate this issue to some extent. Given that software is widely available to apply these techniques on standard survival models, but less common place for msm’s, copulas and frailty models, this is another potential advantage of the dual-outcome approach. Further research is needed to understand the performance of all these methods when penalisation and shrinkage is applied.

6 | CONCLUSIONS

This is the first study to compare modeling techniques for the prediction of the risk of two survival outcomes both occurring in the presence of residual confounding. In the clinical example, the product method resulted in a miscalibrated estimate of the risk of CVD and T2D, indicating the conditional independence assumption was violated despite conditioning on 8 major risk factors. This motivates the need for techniques which appropriately model the residual dependence. In the simulation, all models resulted in similar levels of discrimination, however variable performance was found with respect to discrimination. The dual-outcome and msm methods were the most robust to model misspecification. Although these methods were also the most prone to overfitting (this was observed at minimum sample sizes according to existing criteria), this is an issue that can be solved by reducing the number of predictor variables or recruiting more individuals. On the contrary, the poor calibration of the other methods induced by model misspecification cannot be dealt with. When sample sizes are very large (such as in the clinical example), we recommend the dual-outcome approach alone, due to its comparative performance to the other methods alongside a substantially lower computational time.

AUTHOR CONTRIBUTIONS

Alexander Pate and Glen P. Martin conceived and designed the study in discussion with Matthew Sperrin, Richard D. Riley, Iain Buchan, and Jamie C. Sergeant. Alexander Pate conducted the analysis and interpreted the results in discussion with all authors. Alexander Pate wrote the initial draft of the article with support from Glen P. Martin, which was then critically reviewed for important intellectual content by all authors. All authors have approved the final version of the article.

ACKNOWLEDGEMENTS

The authors would like to thank the Research IT team for their assistance and the use of the Computational Shared Facility at The University of Manchester, on which all the simulations were run.

FUNDING INFORMATION

This work was supported by funding from the MRC-NIHR Methodology Research Programme [grant number: MR/T025085/1].

CONFLICT OF INTEREST STATEMENT

No competing interest.

DATA AVAILABILITY STATEMENT

Reusable code is available from our GitHub public repository.⁷⁶ Data for the clinical example cannot be shared and must be obtained through an application to the Clinical Practice Research Datalink. The simulation was implemented in R version 4.1.2,³² and rstudio¹⁰² using the following packages: mstate,^{63,64} GJRM,^{52,53} rstan,⁷⁸ cubature,¹⁰³ survAUC,¹⁰⁴ mfp,⁹⁸ dplyr,¹⁰⁵ Hmisc,¹⁰⁶ rms,¹⁰⁷ ggplot2,¹⁰⁸ Cairo,¹⁰⁹ DescTools,¹¹⁰ ggpubr,¹¹¹ knitr,¹¹² reshape2,¹¹³ mice,⁸⁹ gems,¹¹⁴ and simsurv.¹¹⁵ Analysis were run on the computation shared facility at University of Manchester.

ETHICS STATEMENT

Access to Clinical Practice Research Datalink data is supported by ISAC protocol 20_000102.


ORCID

Alexander Pate  <https://orcid.org/0000-0002-0849-3458>


Matthew Sperrin  <https://orcid.org/0000-0002-5351-9960>


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

Jamie C. Sergeant  <https://orcid.org/0000-0002-9000-4413>

Tjeerd Van Staa  <https://orcid.org/0000-0001-9363-742X>

Niels Peek  <https://orcid.org/0000-0002-6393-9969>

Mamas A. Mamas  <https://orcid.org/0000-0001-9241-8890>

Martin O'Flaherty  <https://orcid.org/0000-0001-8944-4131>

Iain Buchan  <https://orcid.org/0000-0003-3392-1650>

Glen P. Martin  <https://orcid.org/0000-0002-3410-9472>

REFERENCES

1. Steyerberg EW. In: Gail M, Jonathan S, Singer B, eds. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. Cham: Springer; 2019. doi:10.1007/978-3-030-16399-0
2. Riley RD, van der Windt D, Croft P, Moons KGM. *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. Oxford, UK: Oxford University Press; 2019.
3. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis vs prognosis. *J Clin Epidemiol*. 2021;132:142-145. doi:10.1016/j.jclinepi.2021.01.009
4. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults. *Chest*. 1991;100(6):1619-1636. doi:10.1378/chest.100.6.1619
5. Jentzer JC, Bennett C, Wiley BM, et al. Predictive value of the sequential organ failure assessment score for mortality in a contemporary cardiac intensive care unit population. *J Am Heart Assoc*. 2018;7(6):e008169. doi:10.1161/JAHA.117.008169
6. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357(3):j2099. doi:10.1136/bmj.j2099
7. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham heart study. *Circulation*. 2008;117(6):743-753. doi:10.1161/CIRCULATIONAHA.107.699579
8. Lim WS, van der Eerden MM, Laing R, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax*. 2003;58(5):377-382. doi:10.1136/thorax.58.5.377
9. McAllister KSL, Ludman PF, Hulme W, et al. A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales. *Int J Cardiol*. 2016;210:125-132. doi:10.1016/j.ijcard.2016.02.085
10. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E. FRAX and the assessment of fracture probability in men and women from the UK. *Osteoporos Int*. 2008;19(4):385-397. doi:10.1007/s00198-007-0543-5
11. Caldas C, Greenberg DC, Kearins O, et al. Erratum to: PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res*. 2010;12:1-10. doi:10.1186/bcr2480
12. Nashef SAM, Roques F, Sharples LD, et al. Euroscore II. *Eur J Cardio-Thoracic Surg*. 2012;41(4):734-745. doi:10.1093/ejcts/ezs043
13. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of atrial fibrillation. *J Am Med Assoc*. 2001;285(22):2864-2870. doi:10.1001/jama.285.22.2864
14. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*. 2010;137(2):263-272. doi:10.1378/chest.09-1584
15. Pisters R, Lane DA, Nieuwlaat R, de Vos CB, Crijns HJGM, Lip GYH. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the euro heart survey. *Chest*. 2010;138(5):1093-1100. doi:10.1378/chest.10-0134
16. Chao TF, Lip GYH, Lin YJ, et al. Incident risk factors and major bleeding in patients with atrial fibrillation treated with Oral anticoagulants: a comparison of baseline, follow-up and Delta HAS-BLED scores with an approach focused on modifiable bleeding risk factors. *Thromb Haemost*. 2018;118(4):768-777. doi:10.1055/s-0038-1636534
17. Chao TF, Lip GYH, Liu CJ, et al. Relationship of aging and incident comorbidities to stroke risk in patients with atrial fibrillation. *J Am Coll Cardiol*. 2018;71(2):122-132. doi:10.1016/j.jacc.2017.10.085
18. Wu X, Baig A, Kasymjanova G, et al. Pattern of local recurrence and distant metastasis in breast cancer by molecular subtype. *Cureus*. 2016;8(12):e924. doi:10.7759/cureus.924
19. Peng L, Hong X, Yuan Q, Lu L, Wang Q, Chen W. Prediction of local recurrence and distant metastasis using radiomics analysis of pretreatment nasopharyngeal [18F]FDG PET/CT images. *Ann Nucl Med*. 2021;35(4):458-468. doi:10.1007/s12149-021-01585-9
20. Faria SC, Devine CE, Rao B, Sagebiel T, Bhosale P. Imaging and staging of endometrial cancer. *Semin Ultrasound CT MRI*. 2019;40(4):287-294. doi:10.1053/j.sult.2019.04.001
21. Avanzo M, Stancanello J, Pirrone G, Sartor G. Radiomics and deep learning in lung cancer. *Strahlentherapie Und Onkol*. 2020;196(10):879-887. doi:10.1007/s00066-020-01625-9
22. Uijen A, van de Lisdonk E. Multimorbidity in primary care: prevalence and trend over the last 20 years. *Eur J Gen Pract*. 2008;14(SUPPL. 1):28-32. doi:10.1080/13814780802436093
23. Dhalwani NN, O'Donovan G, Zaccardi F, et al. Long terms trends of multimorbidity and association with physical activity in older English population. *Int J Behav Nutr Phys Act*. 2016;13(1):1-9. doi:10.1186/s12966-016-0330-9
24. van Oostrom SH, Gijzen R, Stirbu I, et al. Time trends in prevalence of chronic diseases and multimorbidity not only due to aging: data from general practices and health surveys. *PLoS One*. 2016;11(8):1-14. doi:10.1371/journal.pone.0160264
25. Hall M, Dondo TB, Yan AT, et al. Multimorbidity and survival for patients with acute myocardial infarction in England and Wales: latent class analysis of a nationwide population-based cohort. *PLoS Med*. 2018;15(3):1-18. doi:10.1371/journal.pmed.1002501
26. Tran J, Norton R, Conrad N, et al. Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the UK between 2000 and 2014: a population-based cohort study. *PLoS Med*. 2018;15(3):1-23. doi:10.1371/journal.pmed.1002513
27. The Academy of Medical Sciences. Multimorbidity: A Priority for Global Health Research; 2018. <https://acmedsci.ac.uk/file-download/82222577>

28. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet*. 2012;380(9836):37-43. doi:[10.1016/S0140-6736\(12\)60240-2](https://doi.org/10.1016/S0140-6736(12)60240-2)
29. Blodgett JM, Rockwood K, Theou O. Changes in the severity and lethality of age-related health deficit accumulation in the USA between 1999 and 2018: a population-based cohort study. *Lancet Heal Longev*. 2021;2(2):e96-e104. doi:[10.1016/S2666-7568\(20\)30059-3](https://doi.org/10.1016/S2666-7568(20)30059-3)
30. European Observatory on Health Systems and Policies, Rijken M, Struckmann V, Dyakova M, Melchiorre MG. ICARE4EU: improving care for people with multiple chronic conditions in Europe. *Eurohealth*. 2013;19(3):29-31.
31. Taylor AW, Price K, Gill TK, et al. Multimorbidity: not just an older person's issue. *BMC Public Health*. 2010;10:718. doi:[10.1186/1471-2458-10-718](https://doi.org/10.1186/1471-2458-10-718)
32. Afshar S, Roderick PJ, Kowal P, Dimitrov BD, Hill AG. Multimorbidity and the inequalities of global ageing: a cross-sectional study of 28 countries using the world health surveys. *BMC Public Health*. 2015;15(1):1-10. doi:[10.1186/s12889-015-2008-7](https://doi.org/10.1186/s12889-015-2008-7)
33. Arokiasamy P, Uttamacharya U, Jain K, et al. The impact of multimorbidity on adult physical and mental health in low- and middle-income countries: what does the study on global ageing and adult health (SAGE) reveal? *BMC Med*. 2015;13(1):1-16. doi:[10.1186/s12916-015-0402-8](https://doi.org/10.1186/s12916-015-0402-8)
34. Garin N, Koyanagi A, Chatterji S, et al. Global multimorbidity patterns: a cross-sectional, population-based, multi-country study. *J Gerontol A*. 2016;71(2):205-214. doi:[10.1093/gerona/glv128](https://doi.org/10.1093/gerona/glv128)
35. Lip GYH, Genaidy A, Tran G, Marroquin P, Estes C, Sloop S. Improving stroke risk prediction in the general population: a comparative assessment of common clinical rules, a new multimorbid index, and machine-learning-based algorithms. *Thromb Haemost*. 2021;122(1):142-150. doi:[10.1055/a-1467-2993](https://doi.org/10.1055/a-1467-2993)
36. Nopp S, Spielvogel C, Schmaldienst S, et al. Bleeding risk assessment in end-stage kidney disease: validation of existing risk scores and evaluation of a machine learning-based approach. *Thromb Haemost*. 2022;122(09):1558-1566. doi:[10.1055/a-1754-7551](https://doi.org/10.1055/a-1754-7551)
37. Pintilie M. *Competing Risks: A Practical Perspective*. Hoboken, NJ: John Wiley & Sons; 2006.
38. Geskus RB. *Data Analysis with Competing Risks and Intermediate States*. Boca Raton: CRC Press; 2016.
39. Beyersmann J, Allignol A, Schumacher M. *Competing Risks and Multistate Models with R*. New York, NY: Springer New York; 2012.
40. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26(11):2389-2430. doi:[10.1002/sim](https://doi.org/10.1002/sim)
41. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*. 2016;133(6):601-609. doi:[10.1161/CIRCULATIONAHA.115.017719](https://doi.org/10.1161/CIRCULATIONAHA.115.017719)
42. Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic models with competing risks. *Epidemiology*. 2009;20(4):555-561. doi:[10.1097/EDE.0b013e3181a39056](https://doi.org/10.1097/EDE.0b013e3181a39056)
43. Govindarajulu US, D'Agostino RB. Review of current advances in survival analysis and frailty models. *Wiley Interdiscip Rev Comput Stat*. 2020;12(6):1-11. doi:[10.1002/wics.1504](https://doi.org/10.1002/wics.1504)
44. Bakoyannis G, Touloumi G. Practical methods for competing risks data: a review. *Stat Methods Med Res*. 2012;21(3):257-272. doi:[10.1177/0962280210394479](https://doi.org/10.1177/0962280210394479)
45. Shihl JH, Louis TA. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*. 1995;51(4):1384-1399.
46. Goethals K, Janssen P, Duchateau L. Frailty models and copulas: similarities and differences. *J Appl Stat*. 2008;35(9):1071-1079. doi:[10.1080/02664760802271389](https://doi.org/10.1080/02664760802271389)
47. Durrleman V, Nikeghbali A, Roncalli T. Which copula is the right one? Tech Rep Groupe Rech Opérationnelle, Crédit Lyon; 2000. doi:[10.2139/ssrn.1123135](https://doi.org/10.2139/ssrn.1123135)
48. Marra G, Radice R. Bivariate copula additive models for location, scale and shape. *Comput Stat Data Anal*. 2017;112:99-113. doi:[10.1016/j.csda.2017.03.004](https://doi.org/10.1016/j.csda.2017.03.004)
49. Emura T, Matsui S, Rondeau V. *Survival Analysis with Correlated Endpoints: Joint Frailty-Copula Models*. Singapore: Springer Singapore; 2019. doi:[10.1007/978-981-13-3516-7](https://doi.org/10.1007/978-981-13-3516-7)
50. Georges P, Arnaud-Guilhem A, Emeric N, Guillaume Q, Thierry R. Multivariate Survival Modelling: A Unified Approach with Copulas; 2001. doi:[10.2139/ssrn.1032559](https://doi.org/10.2139/ssrn.1032559)
51. Nelsen RB. *An Introduction to Copulas*. 2nd ed. New York: Springer Publishing Company; 2006.
52. Marra G, Radice R. GJRM: Generalised Joint Regression Modelling; 2017. <https://cran.r-project.org/web/packages/GJRM/index.html>
53. Marra G, Radice R. Joint regression modeling framework for analyzing bivariate binary data in R. *Depend Model*. 2017;5(1):268-294. doi:[10.1515/demo-2017-0016](https://doi.org/10.1515/demo-2017-0016)
54. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*. 1979;16(3):439-454.
55. Duchateau L, Janssen P. *The Frailty Model*. New York: Springer; 2008.
56. Wienke A. *Frailty Models in Survival Analysis*. Chapman and Hall/CRC Biostatistics Series. Boca Raton: CRC Press; 2011.
57. Balan TA, Putter H. A tutorial on frailty models. *Stat Methods Med Res*. 2020;29(11):3424-3454. doi:[10.1177/0962280220921889](https://doi.org/10.1177/0962280220921889)
58. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand J Stat*. 1978;5(3):141-150.
59. Putter H, Spitoni C. Non-parametric estimation of transition probabilities in non-Markov multi-state models: the landmark Aalen-Johansen estimator. *Stat Methods Med Res*. 2018;27(7):2081-2092. doi:[10.1177/0962280216674497](https://doi.org/10.1177/0962280216674497)
60. Titman AC. Transition probability estimates for non-Markov multi-state models. *Biometrics*. 2015;71(4):1034-1041. doi:[10.1111/biom.12349](https://doi.org/10.1111/biom.12349)

61. de Uña-Álvarez J, Meira-Machado L. Nonparametric estimation of transition probabilities in the non-Markov illness-death model: a comparative study. *Biometrics*. 2015;71(2):364-375. doi:10.1111/biom.12288
62. Pepe MS. Inference for events with dependent risks in multiple endpoint studies. *J Am Stat Assoc*. 1991;86(415):770-778. doi:10.1080/01621459.1991.10475108
63. de Wreede LC, Fiocco M, Putter H. The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Comput Methods Programs Biomed*. 2010;99(3):261-274. doi:10.1016/j.cmpb.2010.01.001
64. de Wreede LC, Fiocco M, Putter H. Mstate: an R package for the analysis of competing risks and multi-state models. *J Stat Softw*. 2011;38(7):1-30.
65. Duchateau L, Janssen P. Extensions of the frailty model. *The Frailty Model*. New York, NY: Springer; 2008. doi:10.1007/978-0-387-72835-3_7
66. Do Ha I, Jeong J-H, Lee YL. *Statistical Modelling of Survival Data with Random Effects: H-Likelihood Approach*. Singapore: Springer; 2017. doi:10.1007/978-981-10-6557-6
67. Rueten-Budde AJ, Putter H, Fiocco M. Investigating hospital heterogeneity with a competing risks frailty model. *Stat Med*. 2019;38(2):269-288. doi:10.1002/sim.8002
68. Katsahian S, Resche-Rigon M, Chevret S, Porcher R. Analysing multicentre competing risks data with a mixed proportional hazards model for the subdistribution. *Stat Med*. 2006;25(24):4267-4278. doi:10.1002/sim.2684
69. Scheike TH, Sun Y, Zhang MJ, Jensen TK. A semiparametric random effects model for multivariate competing risks data. *Biometrika*. 2010;97(1):133-145. doi:10.1093/biomet/asp082
70. Emura T, Shih J-H, Il DH, Wilke RA. Comparison of the marginal hazard model and the sub-distribution hazard model for competing risks under an assumed copula. *Stat Methods Med Res*. 2020;29(8):2307-2327. doi:10.1177/0962280219892295
71. Emura T, Chen Y-H. Gene selection for survival data under dependent censoring: a copula-based approach. *Stat Methods Med Res*. 2016;25(6):2840-2857. doi:10.1177/0962280214533378
72. Zhu H, Lan Y, Ning J, Shen Y. Semiparametric copula-based regression modeling of semi-competing risks data. *Commun Stat Theory Methods*. 2021;51(22):7830-7845. doi:10.1080/03610926.2021.1881122
73. Zhou R, Zhu H, Bondy M, Ning J. Semiparametric model for semi-competing risks data with application to breast cancer study. *Lifetime Data Anal*. 2016;22(3):456-471. doi:10.1007/s10985-015-9344-x
74. Lo S, Wilke RA. A copula model for dependent competing risks. *J R Stat Soc Ser C Appl Stat*. 2010;59(2):359-376. doi:10.1111/j.1467‐9876.2009.00695.x
75. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102. doi:10.1002/sim.8086
76. Pate A, GitHub repository. Manchester Predictive Healthcare Group. MRC-Multi-Outcome-Project-4-Joint-Risk-Prediction-Two-Survival-Processes; 2022. [https://github.com/manchester-predictive-healthcare-group/CHI-MRC-multi-outcome/tree/main/Project %204%20Joint%20Risk%20Prediction%20Two%20Survival%20Processes](https://github.com/manchester-predictive-healthcare-group/CHI-MRC-multi-outcome/tree/main/Project%204%20Joint%20Risk%20Prediction%20Two%20Survival%20Processes)
77. Riley RD, Snell KIE, 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. doi:10.1002/sim.7992
78. Stan Development Team. "RStan: The R Interface to Stan." R Package Version 2.21.3. 2021. <https://mc-stan.org/>
79. van Calster B, Nieboer D, Vergouwe Y, de Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176. doi:10.1016/j.jclinepi.2015.12.005
80. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361-387. doi:10.1002/(SICI)1097-0258(19960229)15:4 < 361::AID-SIM168 > 3.0.CO;2-4
81. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691-698. doi:10.1136/heartjnl-2011-301247
82. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19(4):453-473. doi:10.1002/(SICI)1097-0258(20000229)19:4 < 453::AID-SIM350 > 3.0.CO;2-5
83. Nichols GA, Brown JB. The impact of cardiovascular disease on medical care costs in subjects with and without type 2 diabetes. *Diabetes Care*. 2002;25(3):482-486. doi:10.2337/diacare.25.3.482
84. Einarson TR, Acs A, Ludwig C, Panton UH. Economic burden of cardiovascular disease in type 2 diabetes: a systematic review. *Value Heal*. 2018;21(7):881-890. doi:10.1016/j.jval.2017.12.019
85. Raghavan S, Vassy JL, Ho YL, et al. Diabetes mellitus-related all-cause and cardiovascular mortality in a national cohort of adults. *J Am Heart Assoc*. 2019;8(4):e011295. doi:10.1161/JAHA.118.011295
86. Leon BM. Diabetes and cardiovascular disease: epidemiology, biological mechanisms, treatment recommendations and future research. *World J Diabetes*. 2015;6(13):1246-1258. doi:10.4239/wjd.v6.i13.1246
87. Clinical Practice Research Datalink (CPRD). Primary care data for public health research. Accessed September 16, 2021 <https://www.cprd.com/primary-care>
88. Wolf A, Dedman D, Campbell J, et al. Data resource profile: clinical practice research datalink (CPRD) aurum. *Int J Epidemiol*. 2019;48(6):1740-1740G. doi:10.1093/ije/dyz034
89. van Buuren S, Groothuis-oudshoorn K. mice: multivariate imputation by chained equations. *J Stat Softw*. 2011;45(3):1-67.
90. Therneau TM. A Package for Survival Analysis in S_. version 2.38; 2015. <https://cran.r-project.org/package=survival>
91. Team SD. RStan: The R Interface to Stan; 2021. <https://mc-stan.org/>

92. 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. doi:10.1002/sim.8570
93. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;33(3):517-535. doi:10.1002/sim.5941
94. Harrell FE. *Regression Modeling Strategies*. Cham: Springer; 2015.
95. Gauthier J, Wu QV, Gooley TA. Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. *Bone Marrow Transplant*. 2020;55(4):675-680. doi:10.1038/s41409-019-0679-x
96. Devlin TF, Weeks BJ. Spline functions for logistic regression modelling. Paper presented at: Proceedings of the 11th Annual SAS Users gr Intl Conference, vol. 4; 1986:646-651. <https://support.sas.com/resources/papers/proceedings-archive/SUGI86/Sugi-11-119DevlinWeeks.pdf>.
97. 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*. 2015;85(4):777-793. doi:10.1080/00949655.2013.845890
98. Benner A. Multivariable fractional polynomials. https://cran.r-project.org/web/packages/mfp/vignettes/mfp_vignette.pdf Accessed July 24, 2018.
99. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *J R Stat Soc Ser A Stat Soc*. 1999;162(1):71-94. doi:10.1111/1467-985X.00122
100. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. *Fam Med Community Heal*. 2020;8(1):1-7. doi:10.1136/fmch-2019-000262
101. Heinze G, Wallisch C, Dunkler D. Variable selection - a review and recommendations for the practicing statistician. *Biometrical J*. 2018;60(3):431-449. doi:10.1002/bimj.201700067
102. RStudio. Integrated Development for R. RStudio Team; 2020. <http://www.rstudio.com/>
103. Narasimhan B, Johnson SG, Hahn T, Bouvier A, Kieu K. Cubature: Adaptive Multivariate Integration over Hypercubes; 2021. <https://cran.r-project.org/package=cubature>
104. Potapov S, Adler W, Schmid M. survAUC: Estimators of Prediction Accuracy for Time-To-Event Data; 2012. <https://cran.r-project.org/package=survAUC>
105. Wickham H, Francois R, Henry L, Muller K. Dplyr: A Grammar of Data Manipulation; 2022.
106. Harrell FE Jr. Hmisc: Harrell Miscellaneous; 2021. <https://cran.r-project.org/package=Hmisc>
107. Harrell FE. R Package: rms; 2022. <https://cran.r-project.org/package=rms>
108. Wickham H. ggplot2: Elegant Graphics for Data Analysis; 2016. <https://ggplot2.tidyverse.org>
109. Urbanek S, Horner J. Cairo: R Graphics Device using Cairo Graphics Library for Creating High-Quality Bitmap (PNG, JPEG, TIFF), Vector (PDF, SVG, PostScript) and Display (X11 and Win32) Output; 2022. <https://cran.r-project.org/package=Cairo>
110. Signorell A, Aho K, Alfons A, et al. DescTools: Tools for Descriptive Statistics; 2021. <https://cran.r-project.org/package=DescTools>
111. Kassambara A. ggpubr: "ggplot2" Based Publication Ready Plots; 2020. <https://cran.r-project.org/package=ggpubr>
112. Xie Y. knitr: A General-Purpose Package for Dynamic Report Generation in R; 2021. <https://rdr.io/github/yihui/knitr/man/knitr-package.html>
113. Wickham H. Reshaping data with the reshape package. *J Stat Softw*. 2007;21(12):1-20.
114. Blaser N, Vizcaya LS, Estill J, et al. Gems: an R package for simulating from disease progression models. *J Stat Softw*. 2015;64(10):1-22.
115. Brilleman SL, Wolfe R, Moreno-Betancur M, Crowther MJ. Simulating survival data using the simsurv R package. *J Stat Softw*. 2021;97(3):1-27. doi:10.18637/jss.v097.i03

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: Pate A, Sperrin M, Riley RD, et al. Developing prediction models to estimate the risk of two survival outcomes both occurring: A comparison of techniques. *Statistics in Medicine*. 2023;1-24. doi: 10.1002/sim.9771