Dynamic Prediction in Mixture Cure Models: A Model-Based Landmarking Approach

Marta Cipriani[1]   Marco Alfò[1]   Mirko Signorelli[2]

[1]*Department of Statistical Sciences, Sapienza University of Rome (IT)*
[2]*Mathematical Institute, Leiden University (NL)*

Abstract

Mixture cure models are widely used in survival analysis when a portion of patients is considered cured and is no longer at risk for the event of interest. In clinical settings, dynamic survival prediction is particularly important to refine prognosis by incorporating updated patient information over time. Landmarking methods have emerged as a flexible approach for this purpose, as they allow to summarize longitudinal covariates up to a given landmark time and to use these summaries in subsequent prediction. For mixture cure models, the only landmarking strategy available in the literature relies on the last observation carried forward (LOCF) method to summarize longitudinal dynamics up to the landmark time. However, LOCF discards most of the longitudinal information, does not correct for measurement error, and may rely on outdated values if observation times are far apart. To overcome these limitations, we propose a sequential approach that integrates model-based landmarking within a mixture cure model. Initially, longitudinal covariates are modeled using (generalized) linear mixed models, from which individual-specific random effects are predicted. The predicted random effects are then incorporated as covariates into a Cox proportional hazards cure model.

We investigated the performance of the proposed approach under different cure fractions, sample sizes, and longitudinal data structures through an extensive simulation study. The results show that the model-based strategy provides more refined predictions compared to LOCF, even when the model is misspecified in favour of the LOCF approach. Finally, we illustrate our method using a real-world dataset on renal transplant patients.

Keywords: mixture cure models, model-based landmarking, dynamic prediction, mixed-effects models

1   Introduction

Dynamic prediction models are crucial tools in clinical research, as they leverage baseline and follow-up data to provide updated survival predictions as new information becomes available. This ability to adapt predictions over time is particularly valuable in personalised medicine, where therapeutic strategies can be tailored to the evolving profile of each patient.

Several approaches have been developed to implement dynamic prediction. The time-dependent Cox model[1], for example, incorporates time-varying covariates directly into the hazard function, offering a natural extension of the classical proportional hazards framework. Joint models provide a more integrated solution by simultaneously modelling the longitudinal

trajectory of covariates and the survival process, typically by including individual-specific random effects that link the two submodels[2,3]. An alternative and flexible approach is landmarking, which builds upon prediction models at pre-specified time points, known as landmark times, using only individuals still at risk at those times[4]. In landmarking, repeated measures of time-dependent covariates observed up to the landmark time are first summarised, either through simple synthetic functions or statistical models, and then used as predictors in survival models alongside baseline covariates.

The way in which longitudinal information is summarised clearly plays a central role in determining the performance of these approaches. Landmarking methods can be broadly categorized into *traditional* approaches and *model-based* approaches. Traditional methods[4] are based on simple strategies, such as those based on the last observation carried forward (LOCF) or the average of previous measurements. In contrast, model-based approaches employ more sophisticated statistical tools to capture the underlying structure of longitudinal data. These include linear mixed-effects models (LMM) and multivariate functional principal component analysis (MFPCA), both of which allow rich summaries of the time-varying covariates. Such model-based summaries have been used in combination with several prediction models, including MFPCA with Cox models[5,6], MFPCA with random survival forests[7], LMMs with random survival forests[8], and LMMs with penalized Cox models[9,10].

As interest in prediction methods increases, it has also become evident that the assumption of homogeneity in survival models is often violated. In such cases, cure models offer a valuable framework by explicitly accounting for a subgroup of people who are effectively *cured* and not at risk of experiencing the event of interest, even with long-term follow-up[11]. The two most established approaches in this framework are the mixture cure model (MCM)[12,13] and the promotion time cure model (PTCM)[14]. The MCM explicitly separates cured and susceptible individuals, allowing for independent modelling of the two components, which makes it highly interpretable and flexible. PTCM, on the other hand, avoids explicit classification and models the cure fraction through a hazard-based approach, offering a smoother representation of the transition between risk and cure, but typically requires stronger assumptions on the baseline hazard structure. Given its interpretability and flexibility for modelling, in the following we will focus on the MCM only.

Extending dynamic prediction to settings with a cure fraction has recently attracted growing interest. Most of the existing proposals - with the exception of Shi and Yin (2017)[15], who combined landmarking and MCM using the LOCF strategy - have been developed in the context of PTCMs[16–21]. The approach of Shi and Yin (2017)[15] represents, to our knowledge, the only attempt to implement landmarking within an MCM framework. It relies on a simple LOCF-based summary, which may fail to capture the full longitudinal evolution of covariates, especially when the last observation and the landmark time are far apart, and it is unable to correct for the measurement error that is often present when biomarkers are used as covariates. In addition, the adopted definition of a time-dependent cure probability raises conceptual and methodological concerns that will be discussed in more detail in Section 2.3. In this work, we propose an alternative model-based landmarking method for dynamic prediction in MCMs to address these issues.

The remainder of this paper is organised as follows. In Section 2, we introduce the data structure, the notation, and the basics on MCM and landmarking. In Section 3 we describe the four main steps of the proposed model-based approach. Section 4 describes the metrics used to evaluate predictive performance. In Section 5, we report the design and results of a large-scale simulation study. Section 6 illustrates the application of the method to a real-world case study. Finally, Section 7 concludes the paper with a discussion of the main findings, limitations, and

directions for future research.

2 Notation and background

2.1 Data structure and notation

We consider a dataset with m individuals. For subject $i = 1, \ldots, m$, let $\tilde{t}_i$ and c_i denote the true event and censoring times, respectively, and let $t_i = \min(\tilde{t}_i, c_i)$ be the observed time with event indicator $\delta_i = 1$ if $t_i = \tilde{t}_i$ and $\delta_i = 0$ otherwise.

In addition, we assume that we have observed p covariates (indexed by $l = 1, \ldots, p$) of two types:

- p_Y *longitudinal covariates* $\mathbf{Y}_i$, measured at multiple time points t_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. We denote by $n = \sum_{i=1}^{m} n_i$ the total number of observations. We allow for a completely unbalanced study design in continuous time, whereby each subject can have a different number of repeated measurements and the measurement times can differ across subjects. The vector with the n_i observations of the l-th longitudinal covariate for subject i will be denoted by

$$\mathbf{y}_{il} = \big(y_{il}(t_{i1}), \ldots, y_{il}(t_{in_i}) \big)'.$$

- $p - p_Y$ *time-fixed covariates*, which we partition into $\mathbf{X}_i$ and $\mathbf{Z}_i$, representing two sets of baseline covariates (possibly overlapping) that will be linked to different components of the MCM.

2.2 Mixture cure models without longitudinal covariates

MCMs assume that the population is divided into *cured* ($G = 0$) and *uncured* ($G = 1$) individuals. Cured individuals will never experience the event of interest (e.g. relapse after treatment), whereas uncured individuals may eventually do so. However, the cure status G is not always observable in practice, and for each individual the model provides an estimate for the probability of being uncured.

This probability, i.e. the *incidence* component, is described by a binary response model, typically a logistic regression:

$$P(G_i = 1 \mid \mathbf{X}_i) = \pi(\mathbf{X}_i) = \frac{\exp\left(\alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i\right)}{1 + \exp\left(\alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i\right)}. \tag{1}$$

For uncured individuals, the survival distribution conditional on covariates $\mathbf{Z}_i$ is referred to as the *latency* component, and it is commonly specified through a Cox proportional hazards (PH) model:

$$S_u(t \mid \mathbf{Z}_i) = \exp\left[-H_0(t)\exp\left(\boldsymbol{\beta}'\mathbf{Z}_i\right)\right], \tag{2}$$

where the subscript u refers to the *uncured* portion of the population and $H_0(t)$ is the baseline cumulative hazard. The overall survival function is then

$$S(t \mid \mathbf{X}_i, \mathbf{Z}_i) = [1 - \pi(\mathbf{X}_i)] + \pi(\mathbf{X}_i)S_u(t \mid \mathbf{Z}_i). \tag{3}$$

Since this model depends on the latent cure status G, the set of parameters is usually estimated using an Expectation-Maximization (EM) algorithm[22,23].

Although mixture cure models are not limited to a PH formulation (see Amico and Van Keilegom (2018)[11] for a review), the Cox PH cure model is a common choice.

2.3 Landmarking in mixture cure models: the LOCF approach

We consider a mixture cure setting in which landmarking is used to incorporate longitudinal information observed up to a pre-specified time t_{ℓ_i}, referred to as the *landmark time*, to predict survival at a future horizon t_{ω_i}. The notation t_{ℓ_i} allows for subject-specific landmark times, as in settings with a natural landmark (e.g., the time of transplantation for patients on a waiting list). In the present work, however, we focus on the simpler case of a common landmark defined in patient time, so that $t_{\ell_i} = t_\ell$ for all individuals.

The only previous attempt to combine landmarking with MCMs is due to Shi and Yin (2017)[15], who proposed a class of landmark cure rate models that incorporate both a cure fraction and time-dependent covariates for dynamic prediction. In their approach, the cure probability is specified through a logistic function depending on time-dependent covariates,

$$P(G = 1 \mid Y(t)) = \pi(Y(t)) = \frac{\exp(\alpha_0 + \alpha' Y(t))}{1 + \exp(\alpha_0 + \alpha' Y(t))}, \tag{4}$$

where the incidence is defined as the limit

$$P(G = 1) = \lim_{t \to \infty} \pi(Y(t)).$$

At a given landmark time t_ℓ, they propose fitting a standard mixture cure model with covariates $Y(t_\ell)$.

While innovative, this construction raises two conceptual issues:

(i) The cure probability is linked to the limiting value $Y(\infty)$ of time-dependent covariates. This implies that the cure status of an individual depends on future covariate values, even in scenarios where the subject could already have experienced the event before those covariates are realised (e.g., death before transplantation in their motivating example). This specification is not coherent with the idea that the prediction at t_ℓ should be based only on the information available up to that time.

(ii) By construction, Shi and Yin (2017)[15] write the cure probability as in (4). However, from a probabilistic point of view, it can be shown that

$$\lim_{t \to \infty} P(G = 1 \mid Y(t)) \neq P(G = 1 \mid Y(t_\ell))$$

and the error associated to such an approximation cannot be controlled. Therefore, their specification may lead to theoretical inconsistency.

To address these limitations, we assume that cure status is determined solely by baseline covariates, while longitudinal covariates influence only the survival distribution of uncured individuals. Specifically, in our case, equations (2) and (3) can be rewritten, respectively, as

$$S_u(t \mid Z_i, Y_i) = \exp\left[-H_0(t) \exp\left(\beta' Z_i + \psi' Y_i(t)\right)\right], \tag{5}$$

and

$$S(t \mid X_i, Z_i, Y_i) = [1 - \pi(X_i)] + \pi(X_i) S_u(t \mid Z_i, Y_i(t)). \tag{6}$$

This ensures that the prediction at the landmark time relies exclusively on the information available up to that point. Furthermore, the cure status is constant over time, and the corresponding probability is defined in a coherent and identifiable way. In fact, it is often clinically reasonable to regard the propensity to cure as fixed at baseline, reflecting inherent patient- or disease-specific characteristics rather than post-baseline evolution.

3 Model-based landmarking in cure models

We propose a model-based landmarking method for MCMs in which a statistical model is used to describe the dynamics of longitudinal covariates up to the landmark time. The summaries derived from these models are then combined with the baseline information for prediction. The procedure follows four main steps, inspired by Signorelli et al. (2021)[10]:

1. model the longitudinal covariates via (G)LMMs;

2. summarize the longitudinal histories through predicted random effects;

3. estimate the MCM using an EM algorithm;

4. generating post-landmark predictions.

We describe these steps in detail below.

3.1 Step 1: modelling the longitudinal covariates

The dynamics in the longitudinal covariates up to the landmark time are described by using either a linear mixed model (LMM)[24] or a generalized linear mixed-effects model (GLMM)[25].

For each longitudinal covariate, we fit a (generalized) LMM. For non-Gaussian responses, we follow the formulation of Achana *et al.*[26], which linearizes the model regardless of the response and provides an alternative implementation of penalized quasi-likelihood in the spirit of Breslow and Clayton[27]. The model is fitted using only measurements collected before the landmark time t_{ℓ_i} and restricted to individuals still at risk at t_{ℓ_i}:

$$Y_{il}(t_{ij}) = g^{-1}\left(\boldsymbol{W}_{il}(t_{ij})\boldsymbol{\gamma}_l + \boldsymbol{V}_{il}(t_{ij})\boldsymbol{b}_{il}\right) + \epsilon_{il}(t_{ij}), \quad t_{ij} \le t_{\ell_i},$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i(t_{\ell_i})$, $l = 1, \ldots, p_Y$, where $n_i(t_{\ell_i})$ denotes the number of measurements available for subject i up to t_{ℓ_i}. Here

- $g(\cdot)$ is a differentiable monotonic link function;

- $\boldsymbol{\gamma}$ is a vector of fixed effects;

- $\boldsymbol{b}_{il} \sim MVN(0, \boldsymbol{\Sigma}_{b_l})$ is a vector of individual-specific random effects, capturing heterogeneity across subjects;

- $\boldsymbol{W}$ is a design matrix with size $m \times p_W$; it usually contains elements of $\boldsymbol{X}, \boldsymbol{Z}$ and polynomials in t;

- $\boldsymbol{V} \subseteq \boldsymbol{W}$ is a design matrix with size $m \times p_V$. Tipically it includes a constant term (corresponding to an individual-specific intercept) and polynomials in t to describe individual-specific dynamics, as in latent growth models[28];

- $\epsilon_{il} \sim MVN(0, \boldsymbol{\Sigma}_{\epsilon_{il}})$ is a Gaussian error term, with $\boldsymbol{\Sigma}_{\epsilon_{il}}$ typically specified as $\sigma_l^2 \boldsymbol{I}_{n_i(t_{\ell_i})}$, assumed to be conditionally independent of $\boldsymbol{b}_{il}$ and uncorrelated with $\boldsymbol{W}_{il}$.

3.2 Step 2: summarizing longitudinal covariates at the landmark time

At the landmark time, each patient's longitudinal history is summarized using the predicted random effects $\hat{\boldsymbol{b}}_{il}$ obtained from the (G)LMM fitted up to t_{ℓ_i}. These random effects provide a concise representation of individual-specific heterogeneity in observed dynamics and can be treated as time-constant predictors in the cure model.

For the LMM case, the predicted random effects $\hat{\boldsymbol{b}}_{il}$ are calculated as follows:

$$\hat{\boldsymbol{b}}_{il} = \mathbb{E}\left(\boldsymbol{b}_{il} \mid \boldsymbol{y}_{il}\right) = \hat{\boldsymbol{\Sigma}}_{b_l} \boldsymbol{V}'_{il} \left[\boldsymbol{V}_{il}\hat{\boldsymbol{\Sigma}}_{b_l}\boldsymbol{V}'_{il} + \hat{\boldsymbol{\Sigma}}_{\epsilon_{il}}\right]^{-1}\left(\boldsymbol{y}_{il} - \boldsymbol{W}_{il}\hat{\boldsymbol{\gamma}}_l\right) \tag{7}$$

that is, the estimator derived using the conditional distribution of $\boldsymbol{b}_{il} \mid \boldsymbol{y}_{il}$ obtained exploiting the joint Gaussianity of $(\boldsymbol{y}_{il}, \boldsymbol{b}_{il})$. The variance components $\hat{\boldsymbol{\Sigma}}_{b_l}$ and $\hat{\boldsymbol{\Sigma}}_{\epsilon_{il}}$ are estimated from the fitted LMM via restricted maximum likelihood (REML).

When the longitudinal outcome does not satisfy the Gaussian assumption, we fit a generalized linear mixed model.
The marginal density of $\boldsymbol{y}_{il}$ is given by

$$f_Y(\boldsymbol{y}_{il}) = \int f_Y(\boldsymbol{y}_{il} \mid \boldsymbol{b}_{il}) f_B(\boldsymbol{b}_{il}) \, d\boldsymbol{b}_{il}, \tag{8}$$

where $f_Y(\boldsymbol{y}_{il} \mid \boldsymbol{b}_{il})$ is the conditional density specified by the regression model as a function of covariates $\boldsymbol{W}_{il}$ and random effects $\boldsymbol{b}_{il}$, and $f_B(\boldsymbol{b}_{il})$ is the prior density for random effects. In both LMMs and GLMMs, if the correlation between observed covariates $\boldsymbol{W}_{il}$ and unobserved heterogeneity $\boldsymbol{b}_{il}$ is an issue, it can be addressed through a Mundlak-type reparameterization[29], adding cluster-level means of time-varying covariates to the fixed effects part of the model.

Since the integral in the marginal density (8) does not have a closed form, a numerical approximation is required. Following Wolfinger and O'Connell[30], we adopt a *penalized quasi-likelihood* (PQL) approach, which relies on a first-order Taylor expansion to linearize the model on the link scale. Specifically, we define the linearized response

$$\boldsymbol{y}^*_{il} = \boldsymbol{W}_{il}\boldsymbol{\gamma}_l + \boldsymbol{V}_{il}\boldsymbol{b}_{il} + \boldsymbol{\epsilon}^*_{il}, \tag{9}$$

where $\boldsymbol{y}^*_{il}$ approximates $\boldsymbol{y}_{il}$ on the link scale for the l-th longitudinal covariate, and

$$\boldsymbol{\epsilon}^*_{il} = [g'(\hat{\boldsymbol{\mu}}_{il})]^{-1}(\boldsymbol{y}_{il} - \hat{\boldsymbol{\mu}}_{il}), \tag{10}$$

with $\hat{\boldsymbol{\mu}}_{il}$ denoting the fitted mean. After linearization, $\boldsymbol{\epsilon}^*_{il}$ is assumed to follow a multivariate Gaussian distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}^*_{il}}$, parametrized analogously to $\boldsymbol{\Sigma}_{\epsilon_{il}}$ in the LMM. The resulting conditional and marginal distributions of $\boldsymbol{y}^*_{il}$ are

$$\boldsymbol{y}^*_{il} \mid \boldsymbol{b}_{il} \sim MVN(\boldsymbol{W}_{il}\boldsymbol{\gamma}_l + \boldsymbol{V}_{il}\boldsymbol{b}_{il}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}^*_{il}}) \tag{11}$$

and

$$\boldsymbol{y}^*_{il} \sim MVN(\boldsymbol{W}_{il}\boldsymbol{\gamma}_l, \boldsymbol{V}_{il}\boldsymbol{\Sigma}_{b_l}\boldsymbol{V}'_{il} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}^*_{il}}). \tag{12}$$

As in the LMM case, the predicted random effects $\hat{\boldsymbol{b}}_{il}$ are then obtained as empirical Bayes estimates from the conditional distribution of $\boldsymbol{b}_{il} \mid \boldsymbol{y}^*_{il}$.

3.3 Step 3: estimation of the cure model parameters

Once summaries for observed time-dependent covariates have been obtained, we consider the set of subjects still at risk at the landmark time and employ a mixture cure model for post-landmark survival prediction, incorporating both baseline covariates (X_i, Z_i) and summaries $\hat{\mathbf{b}}_i = (\hat{b}_{i1}, \dots, \hat{b}_{ip_Y})$. Specifically, while the incidence component is as in (1), the latency in (2) becomes

$$S_u(t \mid Z_i, \hat{b}_i) = \exp[-H_0(t) \exp(\boldsymbol{\beta}' Z_i + \boldsymbol{\psi}' \hat{b}_i)].$$

If we denote by $\boldsymbol{\theta}$ the *complete* set of parameters of interest, the complete-data likelihood function for the Cox PH cure model can be written as

$$\mathcal{L}_c(\boldsymbol{\theta}) = \prod_{i=1}^{n} [\pi(X_i) f_u(T_i \mid Z_i, b_i)]^{\delta_i G_i} [\pi(X_i) S_u(T_i \mid Z_i, b_i)]^{(1-\delta_i) G_i}$$
$$[1 - \pi(X_i)]^{(1-\delta_i)(1-G_i)} \tag{13}$$

In the E-step of the EM algorithm[31], we calculate the posterior expectation of the latent indicator G_i. For individual $i \in (1, \dots m)$ at iteration r, this is given by

$$q_i^{(r)} = \delta_i + (1 - \delta_i) \frac{\pi^{(r-1)}(X_i) S_u^{(r-1)}(T_i \mid Z_i, \hat{b}_i)}{\left[1 - \pi^{(r-1)}(X_i)\right] + \pi^{(r-1)}(X_i) S_u^{(r-1)}(T_i \mid Z_i, \hat{b}_i)} \tag{14}$$

The M-step maximizes the expected log-likelihood replacing G_i by q_i and optimizing over the parameters set.

The *loss* function used in the M-step is the negative of the conditional expectation of the complete data log-likelihood given the observed data and the current parameter estimates $\boldsymbol{\theta}^{(r)}$. This loss function can be decomposed into the sum of two loss sub-functions, one for the incidence and the other for the latency components. That is,

$$\mathbb{E}\left\{\log\left[\mathcal{L}_c(\cdot) \mid \text{data}, \boldsymbol{\theta}^{(r)}\right]\right\} = Q_1\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}\right) + Q_2\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}\right)$$

where the expected log-likelihood contribution for the incidence component can be written as

$$Q_1\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}\right) = \sum_{i=1}^{n} q_i^{(r)} \log \pi(X_i) + (1 - q_i^{(r)})(1 - \delta_i) \log(1 - \pi(X_i)) \tag{15}$$

and $(1 - q_i^{(r)})(1 - \delta_i) = (1 - \delta_i)$ by definition. The contribution for the latency component is

$$Q_2\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}\right) = \sum_{i=1}^{n} q_i^{(r)} \delta_i \log f_u(T_i \mid Z_i, \hat{b}_i) + q_i^{(r)}(1 - \delta_i) \log S_u(T_i \mid Z_i, \hat{b}_i) \tag{16}$$

Maximization with respect to H_0 at iteration r leads to the following modified Breslow estimator[32] of the baseline cumulative hazard function

$$\hat{H}_0^{(r)}(t) = \sum_{i=1}^{m} \frac{I(t_i < t)\delta_i}{\sum_{i \in R_t} q_i^{(r)} \exp\left(\boldsymbol{\beta}^{(r-1)'} Z_i + \boldsymbol{\psi}^{(r-1)'} \hat{b}_i\right)} \tag{17}$$

where the numerator is the number of events up to time t and R_t represents the set of subjects that are still at risk at time t.

3.4 Step 4: post-landmark prediction

Once the regression parameters have been estimated, we can derive the conditional cure and survival probabilities by plugging in the estimated values:

$$\hat{\pi}(X_i \mid t_{\ell_i}) = \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}' X_i)}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}' X_i)}, \tag{18}$$

$$\hat{S}u(t_{\omega_i} \mid Z_i, \hat{b}_i, t_{\ell_i}) = \exp\left(-\hat{H}_0(t_{\omega_i}), \exp(\hat{\beta}' Z_i + \hat{\psi}' \hat{b}_i)\right). \tag{19}$$

Predictions can be obtained for both subjects used in model estimation and for out-of-sample individuals, since they rely on the parameter estimates obtained in Step 3. This naturally raises the question of how to evaluate the quality of such predictions. The next section describes the accuracy measures that we suggest employing for this purpose.

4 Metrics for predictive performance evaluation

In the MCM framework, predictive performance is typically assessed separately[33] for the *incidence* and the *latency* component.

For the incidence component, since the true cure status G_i is only partially observed, we rely on weighted versions of standard classification metrics, where each subject is weighted by its estimated posterior probability of being uncured $\hat{q}_i$. Discrimination is quantified by the *weighted AUC*[34],

$$\widehat{\text{AUC}}_{inc} = \frac{\sum_{i \neq j} I(\hat{\eta}_i^{(I)} > \hat{\eta}_j^{(I)}) \, \hat{q}_i \, (1 - \hat{q}_j)}{\sum_{i \neq j} \hat{q}_i \, (1 - \hat{q}_j)},$$

where $\hat{\eta}_i^{(I)} = \hat{\alpha}_0 + \hat{\alpha}' X_i$ is the incidence linear predictor and $\hat{q}$ is the posterior expectation in (14) obtained at the convergence of the EM algorithm. Prediction error is measured by the *weighted Brier score*[35],

$$\widehat{\text{BS}}_{inc} = \frac{1}{m} \sum_{i=1}^{m} \hat{u}_i \, (\hat{q}_i - \tilde{G}_i)^2,$$

where $\tilde{G}_i = 1$ if the subject i experienced the event, $\tilde{G}_i = 0$ if censored beyond the largest observed event time, and $\hat{u}_i$ are the inverse probability of censoring weights (IPCW), i.e. the inverse of the Kaplan–Meier estimate of the probability of remaining uncensored up to the observed time.

For the latency component, prediction performance is evaluated using standard survival measures modified for censoring. Discrimination at time t is assessed by the *time-dependent AUC*[36],

$$\widehat{\text{AUC}}(t)_{lat} = \frac{\sum_{i \neq j} I(\hat{\eta}_i^{(L)} > \hat{\eta}_j^{(L)}) \, \delta_i(t) \, (1 - \delta_j(t)) \, \hat{u}_i(t) \, \hat{u}_j(t)}{\sum_{i \neq j} \delta_i(t) \, (1 - \delta_j(t)) \, \hat{u}_i(t) \, \hat{u}_j(t)},$$

where $\hat{\eta}_i^{(L)} = \hat{\beta}' Z_i + \hat{\psi}' \hat{b}_i$ is the latency linear predictor and $\delta_i(t)$ indicates whether subject i experienced the event before t. Prediction error at time t is measured by the *time-dependent Brier score*[37],

$$\widehat{\text{BS}}(t)_{lat} = \frac{1}{m} \sum_{i=1}^{m} \hat{u}_i(t) \left(I(t_i > t) - \hat{S}_i(t)\right)^2,$$

where $\hat{S}_i(t)$ is the predicted survival probability for subject i and $\hat{u}_i(t)$ are the corresponding IPCW weights. Finally, the overall discriminative ability can be summarized by the *concordance index (C-index)*[38,39],

$$\widehat{C} = \frac{\sum_{i \neq j} I(\hat{\eta}_i^{(L)} > \hat{\eta}_j^{(L)}) D_{i,j}}{\sum_{i \neq j} D_{i,j}},$$

with $D_{i,j} = I[t_i < t_j, \delta_i = 1] + I[t_i = t_j, \delta_i = 1, \delta_j = 0]$.

5 Simulation study

5.1 Simulation design

We conducted a simulation study to evaluate the performance of the proposed model-based approach for dynamic prediction in mixture cure models, comparing it with a LOCF strategy. The study considered different sample sizes, cure fractions, number of repeated measurements, and data-generating mechanisms.

For each setting, 1000 datasets including four time-invariant and four time-varying covariates were generated. Subjects were followed longitudinally up to a pre-specified landmark time ($t_\ell = 3$), with M_i repeated measurements of the time-varying covariates. The time-varying covariates were simulated as linear trajectories with subject-specific random intercepts and slopes drawn from normal distributions, $N(0, 1)$ and $N(0, 0.7)$ respectively, and standard Gaussian measurement-level noise was added, $\varepsilon_{il} \sim N(0, 1)$. Two longitudinal designs were considered: a *balanced* design with $M_i = 10$ measurements per subject, and an *unbalanced* design obtained by randomly deleting some of the observation times so that $M_i \in [5, 10]$. Under the unbalanced design, the LOCF strategy, applied after the amputation step, relied on the last observed measurement after deletion, which could occur well before the landmark. From the landmark onwards, all subjects were considered at risk, with event times simulated from a Weibull distribution conditional on the time-varying covariates and regression coefficients $\boldsymbol{\beta} = (1, 0, -1, 0)$. Independent censoring times were drawn from an exponential distribution starting at the landmark.

To assess the robustness of the proposed model to misspecification, three data-generating mechanisms were considered for the latency component: (i) a *strongly misspecified* setting, where time-varying covariates and survival times were linked using $\boldsymbol{Y}(t)$; (ii) a *mildly misspecified* setting, where the hazard model in the latency component depends on the current value of the longitudinal process $\boldsymbol{Y}(t_\ell)$; and (iii) a *correctly specified model*, where event times are generated using the random effects approach underlying the proposed method. For each of the 1000 simulated datasets, predictions were obtained on an external validation dataset, generated independently but under the same data-generating mechanism as the corresponding training dataset.

Time-invariant covariates are simulated from a multivariate normal distribution, $N_4(\mathbf{0}, \mathbf{I})$. The cure status was drawn from a Bernoulli distribution via a logistic model for the probability of *not* being cured, conditional on time-invariant covariates. To vary the cure fraction, only the intercept was modified, while the remaining coefficients were fixed at $(-1, 0, 1, 0)$. Specifically, an intercept value $\alpha_0 = 2$ yielded, on average, a cure fraction of 20%, whereas $\alpha_0 = 0.65$ yielded 40%.

Each combination of cure fraction and longitudinal design (balanced or unbalanced) was tested at two sample sizes ($m = 300, 1000$), resulting in four basic scenarios replicated twice.

9

An overview of all simulation scenarios is provided in Table 1. All code and materials to reproduce the simulation study are available in a dedicated GitHub repository at `https://github.com/martacip/cure_prediction`.

5.2 Simulation results

The results for the incidence component are reported in Table 2. Since both LOCF and the model-based approach rely on the same specification for the cure model, their predictive performance in terms of incidence is essentially indistinguishable across scenarios. Small fluctuations are observed, but these appear to be random and unrelated to the sample size or design and do not seem to indicate any systematic advantage of either method.

Figures 1, 2, and 3 show the time-dependent AUC for latency prediction under a strong misspecified, a mild misspecified, and a correctly specified model, respectively.

Under strong misspecification (Figure 1) and a balanced longitudinal study, the LOCF approach outperformed the model-based strategy. This finding is expected because, in this specific design, LOCF coincides with the true data-generating mechanism. However, when the longitudinal design was unbalanced, the proposed model-based approach outperformed the LOCF approach even though survival times were generated from the LOCF mechanism. This is explained by the fact that, due to irregular deletion, the LOCF estimation relied on measurements that could be substantially earlier than the landmark, effectively making the fitted LOCF model itself misspecified. This led to a marked loss of predictive performance, in line with the well-documented limitations of LOCF-based approaches under irregular observation patterns[40,41].

When moving to a mildly misspecified setting (Figure 2) or to a correctly specified setting (Figure 3), the proposed model-based method consistently outperformed LOCF across all scenarios. This reflects the ability of the random-effects formulation to correctly capture individual trajectories when the data-generating process is at least approximately aligned with the assumed structure. Importantly, these conclusions were valid for both $n = 300$ and $n = 1000$, indicating that the comparative performance of the methods was not driven by the sample size. Additional performance metrics, reported in Figures 1-3 of the Supporting Information material, confirmed the patterns observed looking at the time-dependent AUC.

6 Case study: renal transplant data

6.1 Data description and statistical modelling

We illustrate the proposed methodology using the `renal` dataset from the `joineRML` R package[42]. This dataset comprises 407 patients with chronic kidney disease who received a primary renal transplant from either a deceased or living donor at the University Hospital of the Catholic University of Leuven (Belgium) between January 1983 and August 2000. Chronic kidney disease is characterized by progressive loss of renal function, typically assessed through repeated measurements of glomerular filtration rate (GFR).

From a clinical perspective, renal transplantation is a potentially curative intervention, as successful transplantation restores kidney functions and long-term survival can be similar to that of the general population[43,44]. Accordingly, the Kaplan–Meier survival curve for this dataset (Figure 4) exhibits a plateau, a hallmark of cure scenarios[11]. In addition, such datasets have been widely used in methodological studies on joint modelling and dynamic prediction for

renal transplant outcomes[45,46], providing a rich framework for demonstrating the performance of advanced prediction techniques for graft failure.

For each patient, multiple longitudinal biomarkers were recorded over time, along with a time-to-event outcome indicating graft failure. We used baseline covariates, age, weight, and sex in the incidence submodel, with age categorized according to quartiles. Besides the marginal effect of age we also considered its interaction with sex. In the latency submodel, only longitudinal biomarkers reflecting kidney function were included, namely haematocrit (% of red blood cells in blood) and GFR (ml/min/1.73 m^2). For these two biomarkers, we report in Table 3 the average number of measurements per individual at landmark times 1, 1.5, 2, 2.5 and 3 years after transplantation.

In the first step of the proposed method, each longitudinal biomarker was modeled using a linear mixed-effects model. To maintain computational simplicity and interpretability at the individual level, each model included only a random intercept and a random slope for time, allowing patient-specific deviations in both baseline levels and trajectories over time. This specification mirrors that used in the simulation study, without incorporating any additional model validation strategy.

6.2   Results

Like in our simulation study, we evaluated the performance of the model-based approach and compared it to the LOCF-based strategy. We used repeated cross validation with 4 folds and 10 repetitions to evaluate predicted performance.

It is important to note that the renal transplant dataset represents an atypical scenario: each patient has a very large number of repeated measurements (Table 3), which can make modeling individual trajectories more challenging, while the abundance of data naturally favors LOCF. Despite these data characteristics, two distinct patterns emerged.

First, the model-based approach shows an increasing relative advantage as the prediction horizon extends further from the fixed landmark time (typically after the first two years), as illustrated by the time-dependent AUC curves in Figure 5. This highlights the benefit of explicitly modeling subject-specific trajectories rather than relying on the last available observation, which quickly loses relevance as time progresses.

Second, when landmarks are set further from the baseline, the model-based approach also tends to perform better, since additional longitudinal information is available to inform the mixed-effects model. We can also notice that later landmarks are associated with fewer events available to estimate survival, leading to greater variability in performance metrics.

Only AUC results are reported here; results based on the time-dependent Brier score, provided in Figure 5 of the Supporting Information material, showed trends consistent with the AUC but with smaller discrepancies between the two methods and reduced variability. In particular, both the evolution across landmark times and the performance decay when moving further from the landmark year appeared more stable with the Brier score. For clarity, Figure 5 displays AUC curves without standard errors, while plots including standard errors for both the AUC and the Brier score are available in the Supporting Information material (Figures 4 and 6).

Table 5 reports the C-index for the latency submodel, providing a global measure of discriminative ability across landmarks. These values confirm the pattern observed with the AUC curves, with the model-based approach generally outperforming the LOCF-based one.

For the incidence submodel, performance (Table 4) is broadly comparable between the model-based and LOCF-based approaches, consistent with the simulation study, reflecting that baseline covariates largely drive event occurrence predictions.

7 Discussion

In this work, we developed a methodology for dynamic prediction in the framework of mixture cure models with longitudinal covariates. Our approach is based on a model-based landmarking strategy in which we employ a mixed-effects model to obtain subject-specific summaries of the trajectories described by the longitudinal biomarkers. The method makes it possible to include longitudinal covariates into the latency component while maintaining computational tractability, making a step further in the literature on cure models, which so far has mainly focused on static covariates and LOCF-based landmarking.

The simulation study provided systematic evidence of the advantages of the model-based approach. Even when the data-generating mechanism coincided with LOCF, the proposed method performed competitively. Furthermore, once the longitudinal design became unbalanced, the proposed strategy showed clear superiority, reflecting the well-known limitations of LOCF-based approaches when measurement times are irregular. Under more realistic data-generating processes, whether mildly misspecified or correctly specified, the model-based approach consistently outperformed the LOCF-based one across all sample sizes, cure fractions, and longitudinal designs. These findings highlight the robustness of the proposed method in modelling misspecifications and its ability to leverage longitudinal information more effectively. Importantly, improvements were observed for both moderate and large sample sizes, suggesting that the benefits are not linked to the sample size.

The renal transplant case study further illustrated the practical relevance of our methodology. This dataset presents an atypical setting: patients undergo frequent follow-up measurements, leading to an abundance of longitudinal data that naturally favors the LOCF-based approach. Nevertheless, our results demonstrated that the model-based approach provides increasing predictive gains as the prediction horizon extends further from the landmark. This reflects its ability to model underlying trajectories rather than relying on a single past measurement, which becomes increasingly uninformative over time. At the same time, we observed the inherent trade-off of landmarking methods: while later landmarks benefit from richer longitudinal histories, they are associated with fewer events, which translates into greater variability of predictive metrics.

In both simulation and real data, a consistent pattern emerged between the incidence and latency components of the cure model. As the incidence component is primarily based on baseline covariates, the prediction performance for this component remained comparable between the two methods. In contrast, for the latency submodel, which directly incorporates longitudinal information, we observed substantial differences, with the model-based strategy showing clear advantages. This confirms the importance of explicitly modeling individual trajectories when the focus is on residual survival among the susceptible population.

From a methodological perspective, our study emphasizes the value of embedding longitudinal processes into cure models for prediction. While the LOCF-based approach remains attractive for its simplicity, it is prone to inaccuracy in settings with irregular follow-up and offers limited flexibility for extrapolating subject-specific trends. The proposed approach provides a principled alternative that strikes a balance between interpretability, computational feasibility, and predictive accuracy.

Despite its advantages, the proposed methodology also has some limitations. A key aspect of the model-based strategy is that it requires the specification of an appropriate mixed-effects model for the longitudinal trajectories in the first step. In practice, selecting the best-fitting structure (e.g., choice of random effects, covariance structure, or functional form of time) may not be straightforward, especially when data exhibit complex or highly irregular patterns.

Misspecification at this stage can propagate to the prediction step, potentially reducing the accuracy of the latency component. Moreover, in some applications, simple linear time trends may be insufficient to capture the underlying biomarker dynamics, and more flexible models such as spline-based formulations or non-linear time effects may be necessary. While these extensions increase modeling accuracy, they also introduce additional challenges in terms of interpretation, computational burden, and risk of overfitting.

Future research may extend this framework in several directions. Incorporating multivariate longitudinal processes, for instance, through latent process mixed models, would allow joint modeling of multiple biomarkers that reflect a common biological pathway. Another avenue is to explore alternative parametrizations of the model, in which the summary of the longitudinal process used in the landmarked cure model is not the predicted random effects, but rather the predicted value of the longitudinal variable at the landmark time obtained from the mixed-effects model.

Finally, we note that in the present work the landmark time was fixed and common to all individuals, defined in patient time. However, in other applications, the landmark time may arise naturally and be subject-specific, such as the time of transplantation for patients on a waiting list, after which survival prediction is of interest. In such settings, landmarking requires careful methodological adjustments because the length of the follow-up prior to the landmark differs across subjects. As shown by Gomon et al. (2023)[6], this may induce a selection bias in the estimation of both fixed and random effects of the longitudinal model, ultimately degrading predictive performance. Developing strategies, for instance, through appropriate reweighting, to account for individual-specific landmarks while preserving consistency of dynamic prediction in cure models represents an important direction for future research.
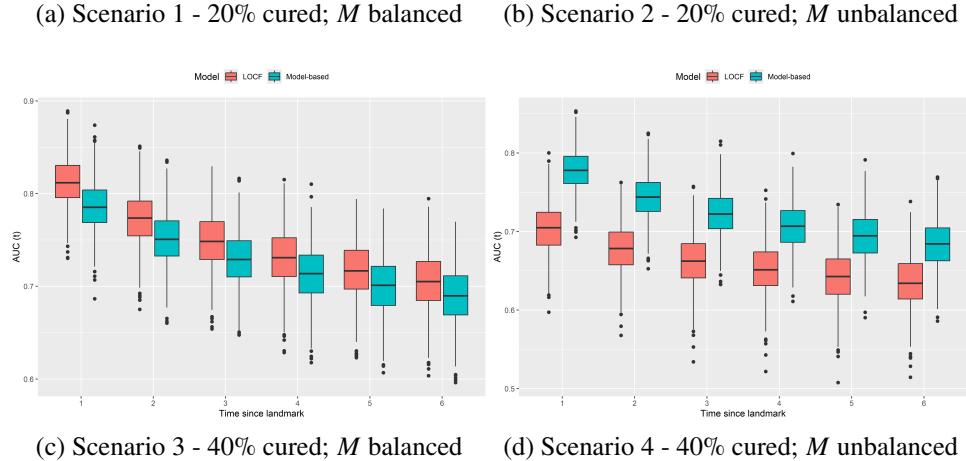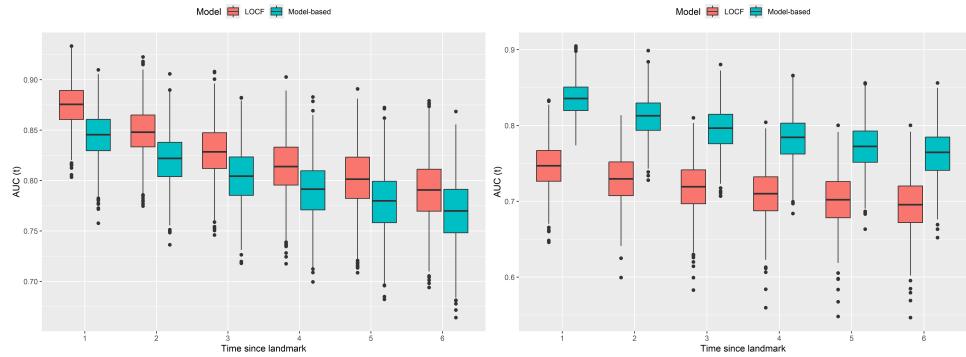
References

[1] Therneau TM, Grambsch PM. *Modeling survival data: extending the Cox model*. Springer New York, NY; 2000. 350 p.

[2] Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics*. 1997;53(1):330-9.

[3] Rizopoulos D. *Joint Modeling of Longitudinal and Time-to-Event Data: Challenges and Future Directions. In: Torelli, N., Pesarin, F., Bar-Hen, A. (eds) Advances in Theoretical and Applied Statistics. Studies in Theoretical and Applied Statistics().* Springer, Berlin, Heidelberg; 2013.

[4] van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*. 2007;34(1):70-85.

[5] Li K, Luo S. Dynamic prediction of Alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Statistics in Medicine*. 2019;38(24):4804-18.

[6] Gomon D, Putter H, Fiocco M, Signorelli M. Dynamic prediction of survival using multivariate functional principal component analysis: A strict landmarking approach. *Statistical Methods in Medical Research*. 2024;33(2):256-72.

[7] Lin J, Li K, Luo S. Functional survival forests for multivariate longitudinal outcomes: Dynamic prediction of Alzheimer's disease progression. *Statistical Methods in Medical Research*. 2021;30(1):99-111.

[8] Devaux A, Proust-Lima C, Geneur R. Random Forests for time-fixed and time-dependent predictors: The DynForest R package. *arXiv:230202670 [statML]*. 2023.

[9] Signorelli M, Spitali P, Al-Khalili Szigyarto C, MARK-MD Consortium, Tsonaka R. Penalized regression calibration: A method for the prediction of survival outcomes using complex longitudinal and high-dimensional data. *Statistics in Medicine*. 2021;40(27):6178-96.

[10] Signorelli M. pencal: an R Package for the Dynamic Prediction of Survival with Many Longitudinal Predictors. The R Journal. 2025;16(2):134-53.

[11] Amico M, van Keilegom I. Cure models in survival analysis. *Annual Review of Statistics and Its Application*. 2018;5:311-42.

[12] Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1949;11(1):15-44.

[13] Berkson J, Gage RP. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*. 1952;47(259):501-15.

[14] Yakovlev AY, Tsodikov AD. *Stochastic models of tumor latency and their biostatistical applications. In: Series in Mathematical Biology and Medicine*. World Scientific; 1996. 269 p.

[15] Shi H, Yin G. Landmark cure rate models with time-dependent covariates. *Statistical Methods in Medical Research*. 2017;26(5):2042-54.

[16] Lambert P, Bremhorst V. Inclusion of time-varying covariates in cure survival models with an application in fertility studies. *Journal of the Royal Statistical Society Series A: Statistics in Society*. 2020;183(1):333-54.

[17] Beretta A, Heuchenne C. Variable selection in proportional hazards cure model with time-varying covariates, application to US bank failures. *Journal of Applied Statistics*. 2018;46(9):1529-49.

[18] Law NJ, Taylor JM, Sandler H. The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*. 2002;3(4):547-63.

[19] Kim S, Zeng D, Li Y, Spiegelman D. Joint modeling of longitudinal and cure-survival data. *Journal of Statistical Theory and Practice*. 2013;7:324-44.

[20] Barbieri A, Legrand C. Joint longitudinal and time-to-event cure models for the assessment of being cured. *Statistical Methods in Medical Research*. 2020;29(4):1256-70.

[21] Xie C, Huang X, Li R, Tsodikov A, Bhalla K. Individual dynamic prediction for cure and survival based on longitudinal biomarkers. *Annals of Applied Statistics*. 2024;18(4):2796-817.

[22] Peng Y, Dear KB. A nonparametric mixture model for cure rate estimation. Biometrics. 2000;56(1):237-43.

[23] Sy JP, Taylor JM. Estimation in a Cox proportional hazards cure model. Biometrics. 2000;56(1):227-36.

[24] McCulloch CE, Searle SR. *Generalized, linear, and mixed models. In: Wiley Series in Probability and Statistics.* John Wiley & Sons, Inc; 2001.

[25] Berridge DM, Crouchley R. *Multivariate generalized linear mixed models using R.* CRC Press; 2011. 304 p.

[26] Achana F, Gallacher D, Oppong R, Kim S, Petrou S, Mason J, et al. Multivariate Generalized Linear Mixed-Effects Models for the Analysis of Clinical Trial–Based Cost-Effectiveness Data. Medical Decision Making. 2021.

[27] Breslow N, DG C. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association. 1993;88(621):9-25.

[28] Bollen KA, Curran PJ. Autoregressive Latent Trajectory (ALT) models a synthesis of two traditions. *Sociological Methods & Research.* 2004;32(3):336-83.

[29] Mundlak Y. On the pooling of time series and cross section data. *Econometrica.* 1978;46(1):69-85.

[30] Wolfinger R, O'Connell M. Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation.* 1993;48(3-4):233-43.

[31] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological).* 1977;39(1):1-38.

[32] Breslow N. Covariates analysis of censored survival data. *Biometrics.* 1974;30(1):89-99.

[33] Asano J, Hirakawa A. Assessing the prediction accuracy of a cure model for censored survival data with long-term survivors: Application to breast cancer data. *Journal of Biopharmaceutical Statistics.* 2017;27(6):918-32.

[34] Zhang Y, Han X, Shao Y. The ROC of Cox proportional hazards cure models with application in cancer studies. *Lifetime Data Analysis.* 2021;27(2):195-215.

[35] Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine.* 1999;18(17-18):2529-45.

[36] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics.* 2000;56(2):337-44.

[37] Schoop R, Graf E, Schumacher M. Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics.* 2008;64(2):603-10.

[38] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA.* 1982;247(18):2543-6.

[39] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*. 2011;30(10):1105-17.

[40] Siddiqui O, Ali MW. A comparison of the random-effects pattern mixture model with last-observation-carried-forward(locf) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics*. 1998;8(4):545-63.

[41] Kenward MG, Molenberghs G. Last Observation Carried Forward: A Crystal Ball? *Journal of Biopharmaceutical Statistics*. 2009;19(5):872-88.

[42] Hickey G, Philipson P, Jorgensen A, Kolamunnage-Dona R, Williamson P, Rizopoulos D, et al. joineRML: Joint Modelling of Multivariate Longitudinal Data and Time-to-Event Outcomes. R package version 0.4.6; 2025.

[43] Wolfe R, Ashby V, Milford E, Ojo A, Ettenger R, Agodoa L, et al. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *The New England Journal of Medicine*. 1999;341(23):1725-30.

[44] Lamb K, Lodhi S, Meier-Kriesche H. Long-term renal allograft survival in the United States: a critical reappraisal. *American Journal of Transplantation*. 2011;11(3):450-62.

[45] Rizopoulos D, Ghosh P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*. 2011;30(12):1366-80.

[46] Fieuws S, Verbeke G, Maes B, Vanrenterghem Y. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics*. 2008;9(3):419-31.

$$m = 300$$

(a) Scenario 1 - 20% cured; *M* balanced   (b) Scenario 2 - 20% cured; *M* unbalanced

(c) Scenario 3 - 40% cured; *M* balanced   (d) Scenario 4 - 40% cured; *M* unbalanced

$$m = 1000$$

(e) Scenario 1 - 20% cured; *M* balanced   (f) Scenario 2 - 20% cured; *M* unbalanced

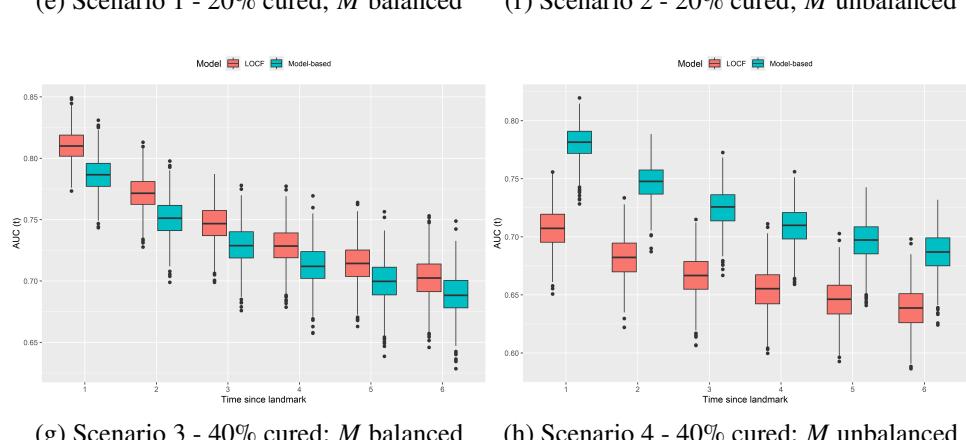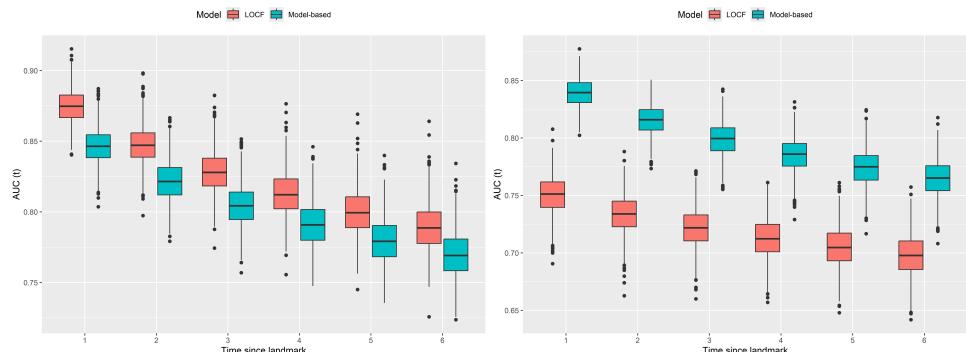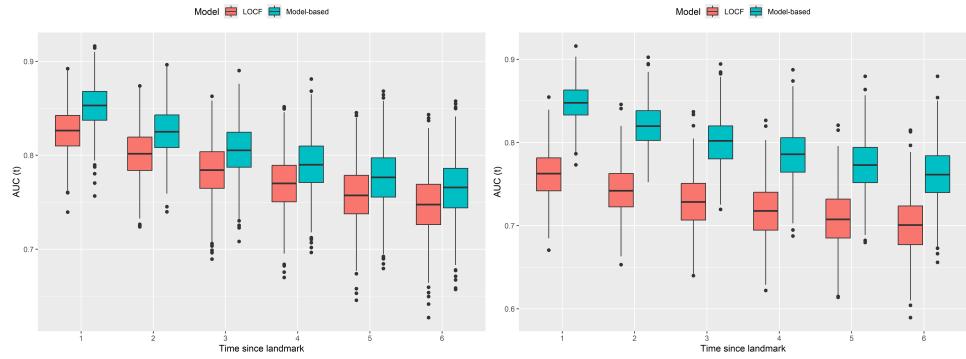(g) Scenario 3 - 40% cured; *M* balanced   (h) Scenario 4 - 40% cured; *M* unbalanced
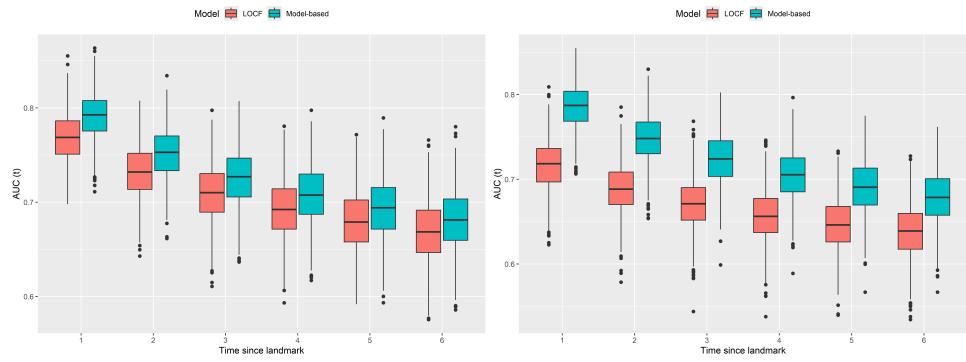
Figure 1: Time-dependent AUC for scenarios under strong misspecification.

$m = 300$

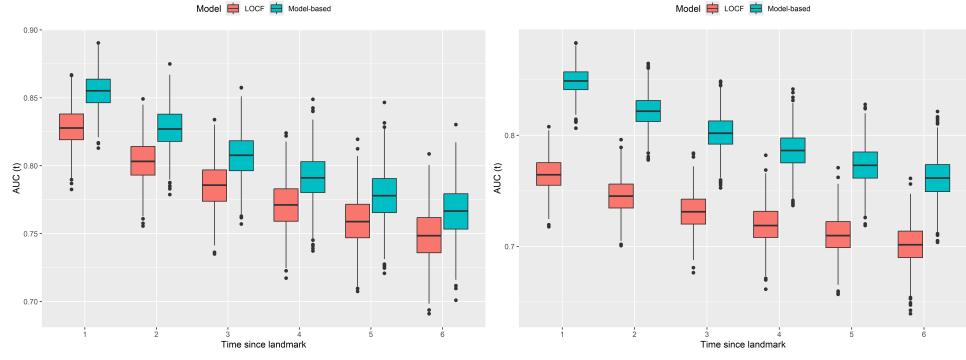(a) Scenario 5 - 20% cured; *M* balanced

(b) Scenario 6 - 20% cured; *M* unbalanced

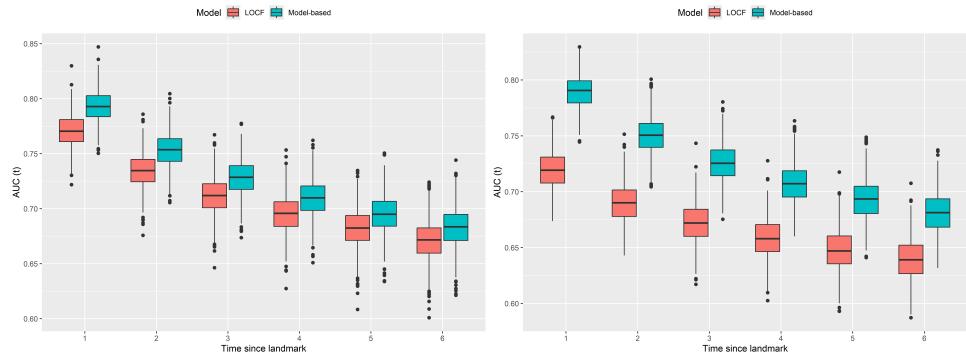(c) Scenario 7 - 40% cured; *M* balanced

(d) Scenario 8 - 40% cured; *M* unbalanced

$m = 1000$

(e) Scenario 5 - 20% cured; *M* balanced

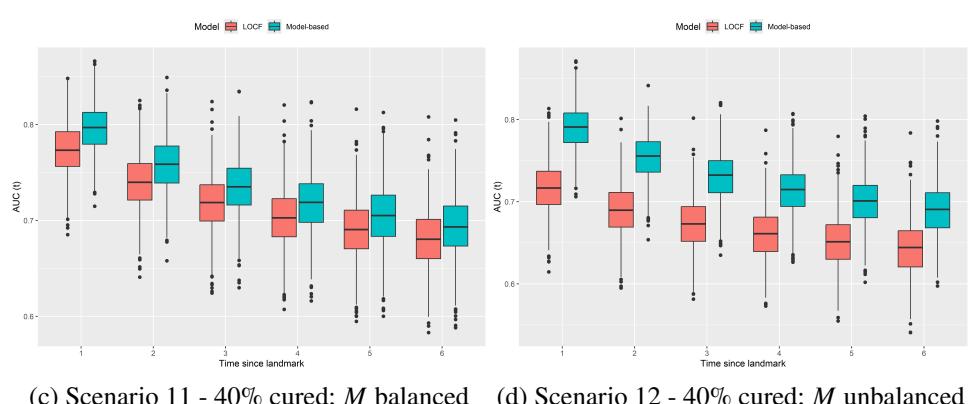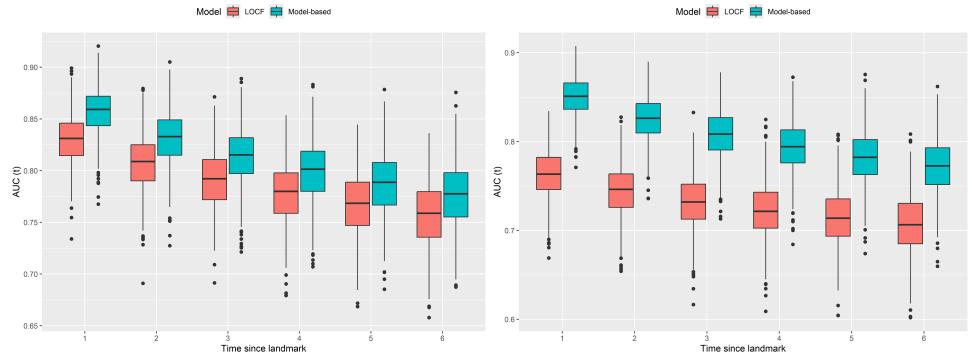(f) Scenario 6 - 20% cured; *M* unbalanced

(g) Scenario 7 - 40% cured; *M* balanced

(h) Scenario 8 - 40% cured; *M* unbalanced

Figure 2: Time-dependent AUC for scenarios under mild misspecification.

$m = 300$

(a) Scenario 9 - 20% cured; M balanced   (b) Scenario 10 - 20% cured; M unbalanced

(c) Scenario 11 - 40% cured; M balanced   (d) Scenario 12 - 40% cured; M unbalanced

$m = 1000$

(e) Scenario 9 - 20% cured; M balanced   (f) Scenario 10 - 20% cured; M unbalanced

(g) Scenario 11 - 40% cured; M balanced   (h) Scenario 12 - 40% cured; M unbalanced
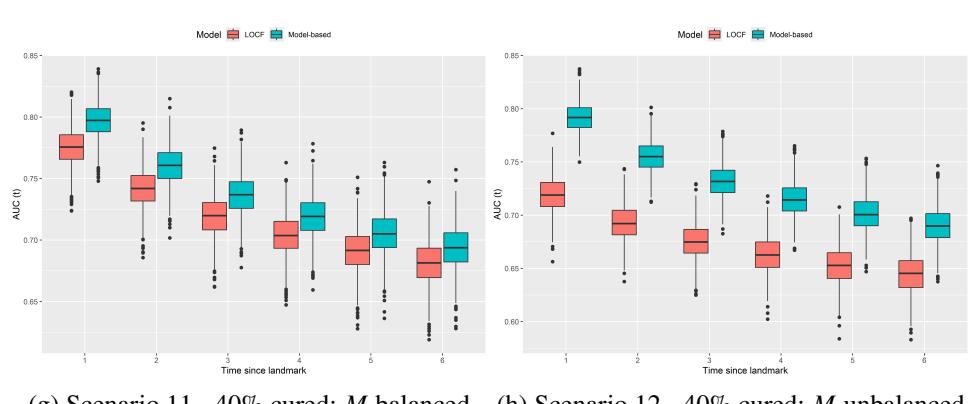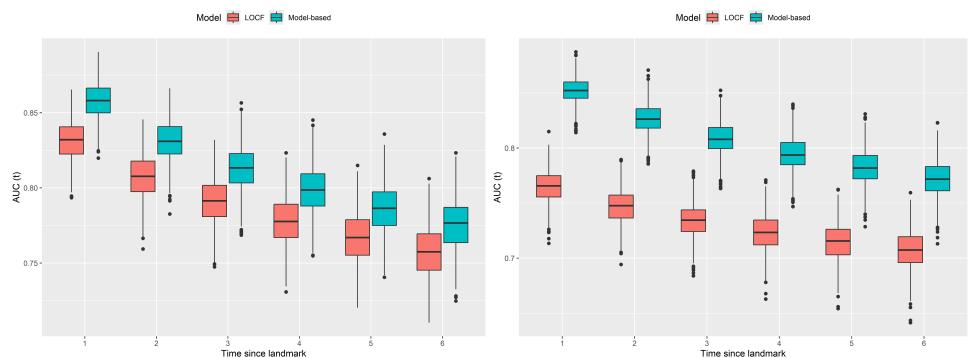
Figure 3: Time-dependent AUC for scenarios under true data-generating process.
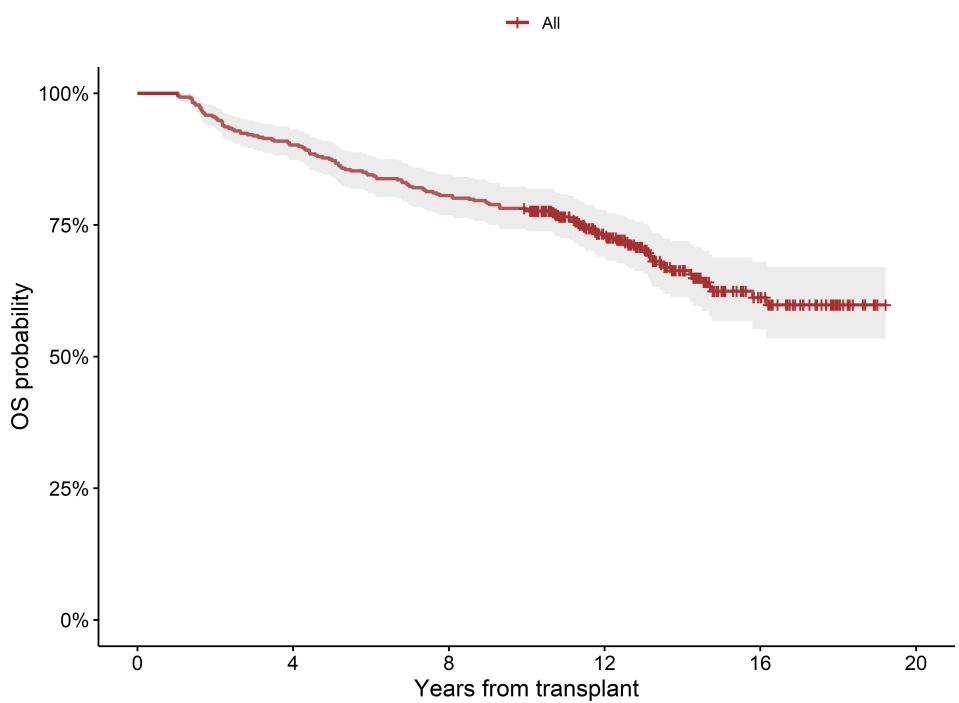
Figure 4: Kaplan-Meier survival curve for renal transplant patients

(a) Landmark time $t = 1$

(b) Landmark time $t = 1.5$

(c) Landmark time $t = 2$

(d) Landmark time $t = 2.5$

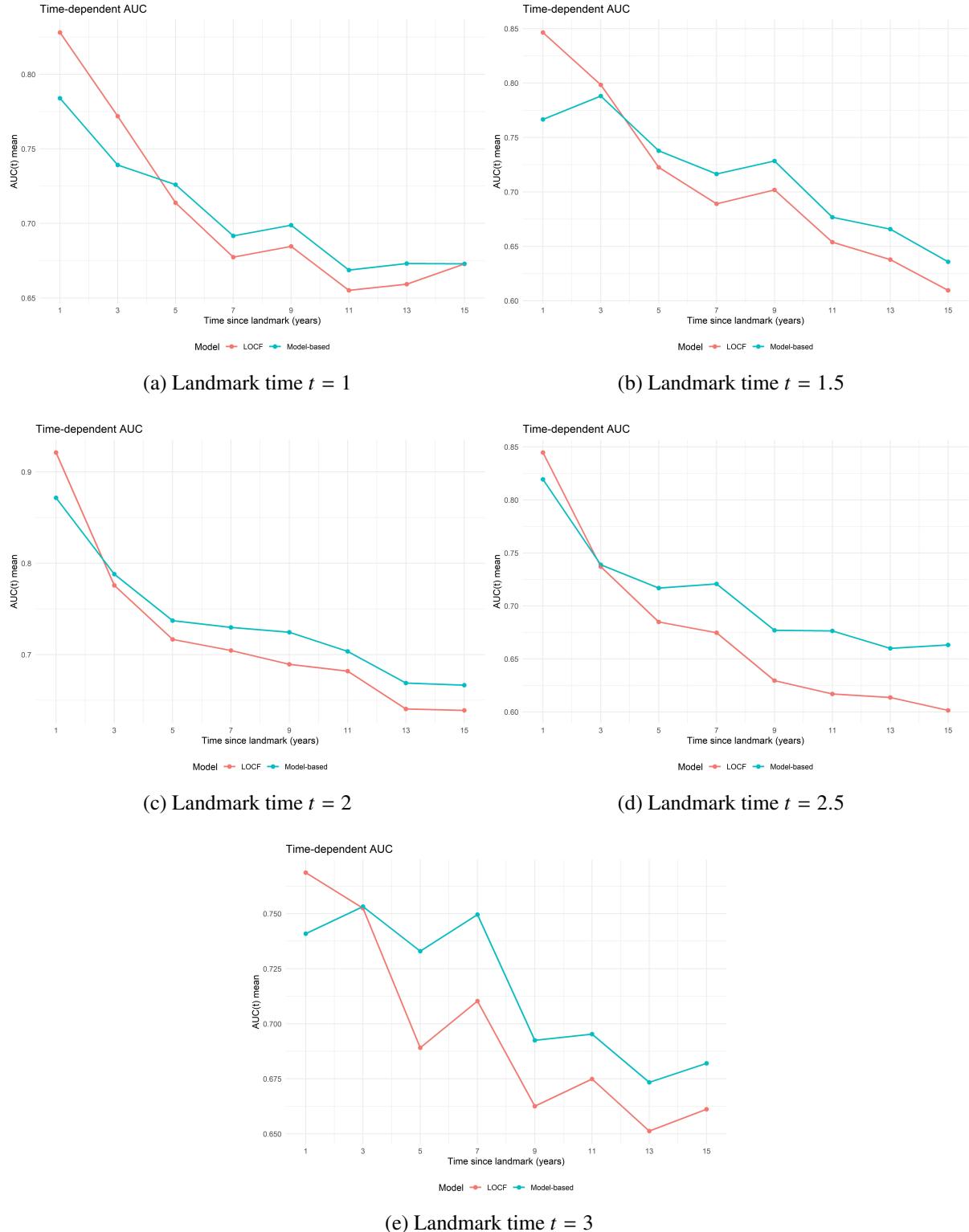(e) Landmark time $t = 3$

Figure 5: Time-dependent AUC curves for the model-based and LOCF approaches at multiple landmark times.

Tables

Table 1: Simulation scenarios with scenario numbers. Each scenario is defined by cure fraction, longitudinal design, and generating mechanism.

Scenario	Cure fraction	Longitudinal design	Generating mechanism
1	20%	Balanced ($M_i = 10$)	Strong misspecification (LOCF)
2	20%	Unbalanced ($M_i \in [5, 10]$)	Strong misspecification (LOCF)
3	40%	Balanced ($M_i = 10$)	Strong misspecification (LOCF)
4	40%	Unbalanced ($M_i \in [5, 10]$)	Strong misspecification (LOCF)
5	20%	Balanced ($M_i = 10$)	Mild misspecification (current value)
6	20%	Unbalanced ($M_i \in [5, 10]$)	Mild misspecification (current value)
7	40%	Balanced ($M_i = 10$)	Mild misspecification (current value)
8	40%	Unbalanced ($M_i \in [5, 10]$)	Mild misspecification (current value)
9	20%	Balanced ($M_i = 10$)	True model (random effects)
10	20%	Unbalanced ($M_i \in [5, 10]$)	True model (random effects)
11	40%	Balanced ($M_i = 10$)	True model (random effects)
12	40%	Unbalanced ($M_i \in [5, 10]$)	True model (random effects)

Table 2: Incidence results for each scenario (mean [SD]) by sample size. Scenario numbers refer to Table 1. Results are reported for LOCF and model-based approaches, with weighted AUC and weighted Brier score.

Scenario	LOCF (AUC [SD]; Brier [SD])	Model-based (AUC [SD]; Brier [SD])
	Sample size n = 300	
1	0.75 [0.04]; 0.14 [0.01]	0.75 [0.04]; 0.15 [0.01]
2	0.75 [0.04]; 0.15 [0.01]	0.75 [0.04]; 0.15 [0.01]
3	0.77 [0.03]; 0.19 [0.01]	0.77 [0.03]; 0.19 [0.01]
4	0.77 [0.03]; 0.19 [0.01]	0.77 [0.03]; 0.19 [0.01]
5	0.75 [0.04]; 0.14 [0.01]	0.75 [0.04]; 0.14 [0.01]
6	0.75 [0.03]; 0.14 [0.01]	0.75 [0.03]; 0.14 [0.01]
7	0.77 [0.03]; 0.19 [0.01]	0.77 [0.03]; 0.19 [0.01]
8	0.77 [0.03]; 0.19 [0.01]	0.77 [0.03]; 0.19 [0.01]
9	0.75 [0.04]; 0.15 [0.01]	0.75 [0.04]; 0.15 [0.01]
10	0.75 [0.04]; 0.15 [0.01]	0.75 [0.04]; 0.14 [0.01]
11	0.77 [0.03]; 0.19 [0.01]	0.77 [0.03]; 0.19 [0.01]
12	0.77 [0.03]; 0.19 [0.01]	0.77 [0.03]; 0.19 [0.01]
	Sample size n = 1000	
1	0.76 [0.02]; 0.14 [0.01]	0.76 [0.02]; 0.14 [0.01]
2	0.76 [0.02]; 0.15 [0.01]	0.76 [0.02]; 0.14 [0.01]
3	0.78 [0.01]; 0.19 [0.01]	0.78 [0.01]; 0.19 [0.01]
4	0.78 [0.01]; 0.19 [0.01]	0.78 [0.01]; 0.19 [0.01]
5	0.76 [0.02]; 0.14 [0.01]	0.76 [0.02]; 0.14 [0.01]
6	0.76 [0.02]; 0.14 [0.01]	0.76 [0.02]; 0.14 [0.01]
7	0.78 [0.01]; 0.19 [0.01]	0.78 [0.01]; 0.19 [0.01]
8	0.78 [0.01]; 0.19 [0.01]	0.78 [0.01]; 0.19 [0.01]
9	0.76 [0.02]; 0.14 [0.01]	0.76 [0.02]; 0.14 [0.01]
10	0.76 [0.02]; 0.15 [0.01]	0.76 [0.02]; 0.14 [0.01]
11	0.77 [0.01]; 0.19 [0.01]	0.77 [0.01]; 0.19 [0.01]
12	0.78 [0.01]; 0.19 [0.01]	0.78 [0.01]; 0.19 [0.01]

Table 3: Average number of longitudinal measurements per patient at selected landmark times

Landmark time (years)	Haematocrit	GFR
1.0	54.07	59.25
1.5	61.55	68.44
2.0	67.31	75.96
2.5	72.69	82.39
3.0	77.56	87.93

Table 4: Incidence metrics (AUC and Brier) at different landmark times for Model-based and LOCF approaches.

Landmark time (years)	Model	AUC Mean	AUC SD	Brier Mean	Brier SD
1	Model-based	0.76	0.03	0.23	0.03
1	LOCF	0.76	0.03	0.23	0.03
1.5	Model-based	0.74	0.03	0.23	0.03
1.5	LOCF	0.75	0.03	0.24	0.03
2	Model-based	0.77	0.03	0.22	0.03
2	LOCF	0.77	0.03	0.23	0.03
2.5	Model-based	0.78	0.03	0.22	0.03
2.5	LOCF	0.79	0.03	0.22	0.04
3	Model-based	0.77	0.03	0.21	0.03
3	LOCF	0.78	0.03	0.22	0.03

Table 5: Latency metrics (C-index) at different landmark times (years) for Model-based and LOCF approaches.

Landmark Time (years)	Model	C-index Mean	C-index SD
1	Model-based	0.66	0.03
1	LOCF	0.66	0.05
1.5	Model-based	0.68	0.05
1.5	LOCF	0.65	0.05
2	Model-based	0.69	0.06
2	LOCF	0.67	0.06
2.5	Model-based	0.68	0.06
2.5	LOCF	0.64	0.07
3	Model-based	0.69	0.06
3	LOCF	0.67	0.06