Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

A copula-based approach for dynamic prediction of survival with a binary time-dependent covariate

Krithika Suresh[1] | Jeremy M.G. Taylor[2] | Alexander Tsodikov[2]

[1]Department of Biostatistics and Informatics, University of Colorado, Aurora, Colorado
[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

Correspondence
Krithika Suresh, Department of Biostatistics and Informatics, University of Colorado, Aurora, CO, 80045, United States Email: krithika.suresh@cuanschutz.edu

Summary

Dynamic prediction methods incorporate longitudinal biomarker information to produce updated, more accurate predictions of conditional survival probability. There are two approaches for obtaining dynamic predictions: (1) a joint model of the longitudinal marker and survival process, and (2) an approximate approach that specifies a model for a specific component of the joint distribution. In the case of a binary marker, an illness-death model is an example of a joint modeling approach that is unified and produces consistent predictions. However, previous literature has shown that approximate approaches, such as landmarking, with additional flexibility can have good predictive performance. One such approach proposes using a Gaussian copula to model the joint distribution of conditional continuous marker and survival distributions. It has the advantage of specifying established, flexible models for the marginals for which goodness-of-fit can be assessed, and has easy estimation that can be implemented in standard software. In this paper, we provide a Gaussian copula approach for dynamic prediction to accommodate a binary marker using a continuous latent variable formulation. We compare the predictive performance of this approach to joint modeling and landmarking using simulations and demonstrate its use for obtaining dynamic predictions in an application to a prostate cancer study.

KEYWORDS:
copula, dynamic prediction, joint analysis, landmark analysis, longitudinal data, survival analysis

1 | INTRODUCTION

Obtaining individualized patient predictions for the risk of a future event is becoming increasingly important in clinical practice. Often survival models are trained using only covariate information collected at a pre-defined clinical time point, such as diagnosis or treatment. However, it is often of interest to obtain predictions at subsequent times and incorporate changing patient information that is collected during follow-up. Dynamic prediction methods use longitudinally collected marker information to produce personalized risk predictions not only at baseline, but also at future time points. There is much literature on developing methods for dynamic prediction, which differ based on the modeling assumptions, structuring of data, and method and computational burden of estimation. The two most common methods for dynamic prediction include joint modeling of the longitudinal and survival data,[1,2,3] and landmarking.[4]

Joint modeling approaches for dynamic prediction involve specifying a model for the longitudinal biomarker (e.g., a linear mixed model), a model for the survival outcome (e.g., Cox proportional hazards) and a method for linking the two (e.g., using

shared random effects).[1,2] This method provides a single, comprehensible model that models the marker process from which we can obtain dynamic predictions for a variety of prediction times. However, it can require restrictive assumptions about the behavior of the marker and survival processes, and computationally intensive techniques for estimation and prediction.

Standard landmarking involves estimating a prediction model at each prediction time point for the sample of subjects who are still at risk at that time.[5] These prediction models are traditionally Cox models, and they incorporate the subject's last available longitudinal information at the prediction time using an imputation method (e.g., last-observation-carried-forward) with administrative censoring applied at the prediction window of interest. Landmarking does not require assumptions about the marker distribution, is easily implementable in standard software, and does not pose a computational burden. However, since it does not provide a comprehensive probability model, predictions are not consistently defined over time and landmarking represents an approximate approach for dynamic prediction.[6] There are several extensions that have been proposed within the landmarking framework that improve prediction but increase computational complexity.[3,7]

In Suresh et al,[8] we propose an approximate approach for dynamic prediction that uses a Gaussian copula to model the joint distribution of a continuous marker and survival time conditional on the prediction time. This method does not constitute a joint model, but allows for predictions to be obtained for any prediction time and window, is not computationally intensive, and provides a greater level of consistency by specifying a single model for the event time distribution. Under several scenarios, we demonstrated that the predictive performance of this method was similar or superior to standard landmarking. This copula method can be thought of as an intermediate approach between landmarking and joint modeling. Joint modeling specifies a stochastic process for the marker, landmarking does not make any distributional assumption about the marker, while the copula approach just specifies the marginal distribution of the marker at each time without explicitly specifying a longitudinal process. Landmarking requires a different survival model at each time of interest, whereas joint modeling and the copula approach each have a single model for the event time distribution.

Much of this presented literature for dynamic prediction focuses on the situation of a continuous marker, whose changing values over time can influence survival. However, during follow-up we may instead collect information on a binary marker that can change during the patient's follow-up, such as an indicator of the occurrence of an intermediate event. In our motivating data set, patients with clinically localized prostate cancer were treated with radiation therapy. During the patient's follow-up, the clinician can detect metastatic clinical failure (binary marker) that can affect the patient's risk of mortality. By incorporating this new information, clinicians can obtain a current, more accurate prediction of a patient's survival to make important medical decisions for the patient, such as additional/modified treatment or increased monitoring frequency.

If the longitudinal marker is a binary variable that can only change from 0 to 1, but not from 1 to 0, then the joint model between the longitudinal marker and the survival outcome can be described by an illness-death model.[9,10] Within the class of multi-state models, under the Markov assumption we can directly obtain the dynamic prediction probabilities by applying the Aalen-Johansen formulas.[11] However, in more realistic and complex situations, obtaining predictions is much more difficult and may require approximation through simulation.[12] In van Houwelingen and Putter,[13] they demonstrate that landmarking methodology can be used as an alternative to multi-state modelling with similar results and easier computation of prediction probabilities. In previous work,[14] we compared the performance of the illness-death model and landmarking with a binary marker under both Markov and semi-Markov assumptions and found that with additional components to make it more flexible, the performance of an approximate approach, such as landmarking, was similar to that of the simple joint model. Thus, based on the advantages provided by the Gaussian copula approach for dynamic prediction with a continuous marker,[8] we explore extending this copula based approach to incorporate a longitudinal binary biomarker.

A Gaussian copula is applicable only when linking two continuous outcomes; however, we are interested in modeling the relationship between a binary marker and the continuous time-to-event outcome. Joint modeling strategies for mixed outcomes using a copula approach were explored by Song et al.[15] We use an extension of their model proposed by de Leon and Wu[16] for mixed polychotomous and continuous outcomes. Using a latent variable formulation of the discrete outcome we transform it into a continuous one, after which we use a Gaussian copula to model the time-varying association between the two continuous outcomes. The advantage of this copula approach is that it allows us to model the marginal distributions of the marker data and time-to-event process and their association separately. This allows us to fit models for the marginals using well-known classes of models and standard goodness-of-fit techniques, and specify a flexible association structure to capture their dependence.

In this paper, we aim to extend a Gaussian copula method for dynamic prediction shown to have good predictive performance and low computational burden to accommodate a longitudinally collected binary marker. In Section 2, we describe the Gaussian copula method for dynamic prediction with mixed outcomes. Using a simulation study, in Section 3 we explore the predictive

performance of our method. We demonstrate the use of our method for our motivating data example of metastatic clinical failure in prostate cancer patients in Section 4. To conclude, in Section 5 we present a discussion and future directions.

2 | METHOD

Consider a survival time distribution T and a marker process $Z(t)$, where T is a continuous outcome and $Z(t)$ is a time-varying marker that is expected to have an influence on the time-to-event outcome. The observed data is given by $\mathcal{D} = \{T_i^*, \Delta_i, \mathbf{Z}_i, \mathbf{X}_i; i = 1, \dots, n\}$, where for individual i, T_i is the true event time, C_i is the censoring time, $T_i^* = \min(T_i, C_i)$ is the observed event time, $\Delta_i = \mathbf{1}(T_i \leq C_i)$ is the censoring indicator, $\mathbf{X}_i$ is the baseline covariate vector, and $\mathbf{Z}_i$ is an $n_i \times 1$ vector observed from the individual's marker process $Z_i(t)$, such that the jth element is given by $z_{ij} = Z_i(\tau_{ij})$ for measurement times $\tau_{ij}, j = 1, \dots, n_i$.

We are interested in obtaining the dynamic prediction of survival for a new individual k from the same population for a prediction window s, conditional on the individual's up-to-date marker information and that the individual has survived up to time τ, which is given by

$$p_k(\tau, s) = \Pr(T_k \geq \tau + s | T_k > \tau, \mathbf{X}_k, \mathbf{Z}_k(\tau)) \tag{1}$$

where $\mathbf{Z}_k(\tau)$ is the history of the marker process for individual k up to time τ, and can be given by the set of longitudinal measurements collected up to time τ or, as we assume in this paper, a scalar of the most recent measurement at time τ, $Z_k(\tau)$.

Since this dynamic prediction is a conditional survival probability that conditions on surviving up to time τ and the marker measurement at time τ, we can instead write it as

$$p_k(\tau, s | \mathbf{X}_k, Z_k(\tau) = z) = \frac{\Pr(T_k \geq \tau + s, Z_k(\tau) = z | T_k > \tau, \mathbf{X}_k)}{\Pr(Z_k(\tau) = z | T_k > \tau, \mathbf{X}_k)} = \frac{\Pr(T_{\tau_k} \geq \tau + s, Z_{\tau_k} = z | \mathbf{X}_k)}{\Pr(Z_{\tau_k} = z | \mathbf{X}_k)}$$

where we define $T_\tau = T | T > \tau$ as the conditional survival time distribution and $Z_\tau = Z(\tau) | T > \tau$ as the cross-sectional marker data at time τ. The subscript τ denotes conditioning on $T > \tau$. Details for this derivation are given in Supplementary Material A. We assume $T_{\tau_i} \sim F_{T_\tau}$ and $Z_{\tau_i} \sim F_{Z_\tau}$ for individual i. F_{T_τ} and F_{Z_τ} are the marginal distributions for the time-to-event outcome and the binary marker data, respectively, conditional on being alive at time τ. Both of these marginals can be conditional on baseline covariates $\mathbf{X}$, which shall be omitted from model specification for brevity. The dynamic prediction is then given by $p(\tau, s) = F_{T_\tau, Z_\tau}(\tau + s, \tau) / F_{Z_\tau}(\tau)$, and we can compute this probability from the marginal distribution F_{Z_τ} and the joint distribution F_{T_τ, Z_τ}. In a joint model, we would specify the full joint distribution of Z and T, and derive the conditional distributions of interest for our prediction. We propose an alternative approximate approach in which we specify marginal distributions for F_{Z_τ} and F_{T_τ} and use a Gaussian copula to give the joint distribution of T_{τ_i} and Z_{τ_i}, from which $p(\tau, s)$ can be obtained.

2.1 | Mixed bivariate copula model and dynamic prediction

Consider our specific situation where the marker process $Z(t)$ is a discrete outcome that can take on only two values at each time τ, i.e., $Z(\tau) = 0$ or 1. Thus, T_τ is continuous and Z_τ is discrete. By Sklar's theorem,[17] a copula is unique if and only if its components are continuous random variables. Thus, we introduce $Z^* \sim F_{Z^*}$, to be an unobserved continuous latent process underlying the discrete marker process Z.[16] The observed Z is related to Z^* through

$$Z(\tau) = \begin{cases} 0, & \text{if } Z^*(\tau) \in (-\infty, 0) \\ 1, & \text{if } Z^*(\tau) \in [0, \infty) \end{cases}$$

We denote $F_{Z_\tau^*}$ as the distribution of $Z_\tau^* = Z^*(\tau) | T > \tau$, i.e., the cross-sectional distribution of Z^* at τ conditional on surviving up to time τ. The joint distribution at τ, F_{T_τ, Z_τ^*}, is then defined by a Gaussian copula as

$$F_{T_\tau, Z_\tau^*}(t, z) = \Phi_2 \left(\Phi^{-1} \left\{ F_{T_\tau}(t) \right\}, \Phi^{-1} \left\{ F_{Z_\tau^*}(z) \right\}; \rho_\tau \right) \tag{2}$$

where Φ is the standard normal distribution, Φ_2 is the standard bivariate normal distribution, and $\rho_\tau = \rho(\tau)$ is the correlation, which is specified as a smooth function of τ and baseline covariates $\mathbf{X}$. In this formulation, the marginals F_{T_τ} and $F_{Z_\tau^*}$ are absolutely continuous distributions. The dynamic prediction of interest at time τ for surviving the prediction window s can then be derived from Eq.(2), the details for which are given in Supplementary Material A. We present separate dynamic prediction formulas conditioning on $Z(\tau) = 0$ and $Z(\tau) = 1$, respectively. In our latent variable formulation, this is equivalent to conditioning on $Z^*(\tau) < 0$, and $Z^*(\tau) \geq 0$, and are given as

$$\Pr(T \geq \tau + s | T > \tau, Z(\tau) = 0) = \Pr(T \geq \tau + s | T > \tau, Z^*(\tau) < 0) \quad (3)$$
$$= \frac{F_{Z^*_\tau}(0) - F_{T_\tau, Z^*_\tau}(\tau + s, 0)}{F_{Z^*_\tau}(0)}$$
$$\Pr(T \geq \tau + s | T > \tau, Z(\tau) = 1) = \Pr(T \geq \tau + s | T > \tau, Z^*(\tau) \geq 0) \quad (4)$$
$$= \frac{[1 - F_{Z^*_\tau}(0)] - F_{T_\tau}(\tau + s) + F_{T_\tau, Z^*_\tau}(\tau + s, 0)}{1 - F_{Z^*_\tau}(0)}$$

2.2 | Copula Components

The components of the copula are specified using flexible, but possibly misspecified, models that aim to provide a good approximation to the true distributions. We select marginal models from well-established survival and regression families for which there are established goodness-of-fit techniques and standard software available. We specify a flexible, parametric form for the association function and use a Gaussian copula due to its tractable nature, allowing us to perform easy estimation with a likelihood-based approach.

2.2.1 | Modeling the binary marker data

For each time τ we specify a simple, flexible model, for the distribution of the marker value where the mean is a function of time τ and baseline covariates $\mathbf{X}$. We can define the latent variable model $Z^*_\tau = \mu(\tau, \mathbf{X}, \boldsymbol{\gamma}) + \epsilon_\tau$ where $\boldsymbol{\gamma}$ is a vector of regression coefficients, $\mu(\tau, \mathbf{X}, \boldsymbol{\gamma})$ is a function of time τ, baseline covariates, and regression coefficients, and ϵ_τ is an error term that is independently, and identically distributed. We do not estimate parameters in the distribution of ϵ_i due to identifiability, so the marginal parameters to be estimated for $F_{Z^*_\tau}$ are given by $\boldsymbol{\theta}_1 = \boldsymbol{\gamma}$. Special examples include,

- If ϵ_τ is normally distributed $N(0, \sigma^2)$, then $Z^*_\tau \sim N(\mu(\tau, \mathbf{X}, \boldsymbol{\gamma}), \sigma^2)$ and Z_τ is a probit model, where $\sigma^2 = 1$ for identifiability.
- If ϵ_τ has a logistic distribution, then Z_τ will be a standard logistic regression.
- If ϵ_τ is non-standardized Student t-distributed $t(0, 1, \upsilon)$ (mean 0, scale 1, and df υ), then $Z^*_\tau \sim t(\mu(\tau, \mathbf{X}, \boldsymbol{\gamma}), 1, \upsilon)$, where we fix unit scale for identifiability.

There are a number of possible data generating models for a longitudinally measured binary marker. If the binary variable can only change from 0 to 1 then we can describe the joint distribution of the marker and survival process with an illness-death model. Under such an illness-death data generating process $Z(t)$ is a binary indicator of the occurrence of an intermediate event prior to the terminal event, and we can write out the distribution of the marker value at τ as

$$\Pr(Z(\tau) = 0 | T > \tau, \mathbf{X}) = \frac{\Pr(Z(\tau) = 0, T > \tau | \mathbf{X})}{\Pr(T > \tau | \mathbf{X})}$$
$$= \frac{e^{-\int_0^\tau \lambda_{01}(u|\mathbf{X}) + \lambda_{02}(u|\mathbf{X})\, du}}{e^{-\int_0^\tau \lambda_{01}(u|\mathbf{X}) + \lambda_{02}(u|\mathbf{X})\, du} + \int_0^\tau e^{-\int_0^\upsilon \lambda_{01}(u|\mathbf{X}) + \lambda_{02}(u|\mathbf{X})\, du} \lambda_{01}(\upsilon|\mathbf{X}) e^{-\int_\upsilon^\tau \lambda_{12}(u|\mathbf{X})\, du}\, d\upsilon}$$
$$\Pr(Z(\tau) = 1 | T > \tau, \mathbf{X}) = 1 - \Pr(Z(\tau) = 0 | T > \tau, \mathbf{X})$$

where $\lambda_{ij}(t|\mathbf{X})$ represents the hazard of transitioning from state i to state j (0: Healthy, 1: Ill, 2: Dead), with transition-specific baseline covariate effects. The details of these derivations are given in Supplementary Material A. Notice that the form of this marginal distribution of $Z(\tau)$ as a function of $\mathbf{X}$ does not correspond to a known distribution. If the true joint distribution between the marker and the survival process are more complex, we can expect that this would also be the case. If the binary variable can change from both 0 to 1 and from 1 to 0, then a possible longitudinal model is a generalized linear mixed model, such as $\text{logit}(\Pr(Z_i(\tau) = 1)) = a_i + b_i\tau + \beta f(\mathbf{X}_i, \tau)$, where a_i and b_i are random effects. Combining this model with a model for the hazard of the event it is feasible to calculate the marginal distribution $\Pr(Z(\tau) = 1 | T > \tau, \mathbf{X})$, but it will also have a complicated functional form as a function of τ and $\mathbf{X}$. Thus, the alternative we described above, using the flexible latent variable model is a misspecified model for the observed marker data that can serve as a good approximation of the true distribution of $Z(\tau)$ at each τ but allows for easy estimation in standard software.

2.2.2 | Modeling the failure time data

We model the time-to-event outcome T using a semiparametric (Cox) or parametric survival model, and consider additional flexibility by allowing for non-proportional hazards or time-varying effects. Thus, we specify the hazard as $h(t) = h_0(t)\exp\{d(t, \mathbf{X}, \boldsymbol{\nu})\}$, where t is time from baseline, $h_0(t)$ is the baseline hazard, $\boldsymbol{\nu}$ is a vector of regression coefficients, and $d(t, \mathbf{X}, \boldsymbol{\nu})$ is a function of baseline covariates, regression coefficients and possibly time to allow for non-proportional hazards and time-varying covariate effects. We note that this model does not include Z. The marginal distribution for the failure time data is then $F_T(t) = 1-\exp\{-\int_0^t h(u)\,\mathrm{d}u\}$. We compute the conditional survival from this model as $F_{T_\tau}(t) = [F_T(t)-F_T(\tau)]/[1-F_T(\tau)]$. Thus, we use a unified single survival model from which we derive the conditional survival distribution at each time τ. The parameters to be estimated are $\boldsymbol{\theta}_2 = \{\boldsymbol{\nu}, H_0(t)\}$, where $H_0(t) = \int_0^t h_0(u)\,\mathrm{d}u$ is the cumulative baseline hazard.

2.2.3 | Modeling the association

Once the marginal models for T_τ and Z^*_τ are specified, we use the Gaussian copula in Eq.(2) to describe the joint distribution between the marker value at τ and failure time process, conditional on surviving up to time τ. The correlation between the marginals is described by the association function ρ_τ, which by definition of the Gaussian copula is restricted to the range $(-1, 1)$. Thus, we reparametrize using Fisher's z-transformation to define $\rho_\tau = [\exp(2\eta_\tau) - 1]/[\exp(2\eta_\tau) + 1]$, where we specify $\eta_\tau = \eta(\tau, \mathbf{X}, \boldsymbol{\theta}_\rho)$ as a function of time τ, baseline covariates $\mathbf{X}$, and association parameters $\boldsymbol{\theta}_\rho$. The association function ρ_τ provides us with information about the magnitude and direction of the correlation between the cross-sectional marker value and the failure time process conditional on being at risk, and whether that relationship changes with time τ or baseline covariates.

2.3 | Estimation

Let $\mathcal{D}$ be the observed data as defined above. Let $\boldsymbol{\theta}$ be the parameter vector containing the respective marginal parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ of F_{T_τ} and $F_{Z^*_\tau}$, and the association parameters $\boldsymbol{\theta}_\rho$. We aim to model the association between the marker and time-to-event processes but consider the correlation due to repeated measurements on the same individual a nuisance. Thus, we assume working independence between measurements taken on each individual at each time and construct a pseudo-likelihood given by

$$\begin{aligned} PL(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} &\Pr(T_{\tau_{ij}} = t_i, Z_{\tau_{ij}} = 0; \boldsymbol{\theta})^{\mathbf{1}(Z(\tau_{ij})=0)\Delta_i} \cdot \Pr(T_{\tau_{ij}} \geq t_i, Z_{\tau_{ij}} = 0; \boldsymbol{\theta})^{\mathbf{1}(Z(\tau_{ij})=0)(1-\Delta_i)} \\ &\cdot \Pr(T_{\tau_{ij}} = t_i, Z_{\tau_{ij}} = 1; \boldsymbol{\theta})^{\mathbf{1}(Z(\tau_{ij})=1)\Delta_i} \cdot \Pr(T_{\tau_{ij}} \geq t_i, Z_{\tau_{ij}} = 1; \boldsymbol{\theta})^{\mathbf{1}(Z(\tau_{ij})=1)(1-\Delta_i)} \\ = \prod_{i=1}^{n}\prod_{j=1}^{n_i} &\Pr(T_{\tau_{ij}} = t_i, Z^*_{\tau_{ij}} < 0; \boldsymbol{\theta})^{\mathbf{1}(Z^*(\tau_{ij})<0)\Delta_i} \cdot \Pr(T_{\tau_{ij}} \geq t_i, Z^*_{\tau_{ij}} < 0; \boldsymbol{\theta})^{\mathbf{1}(Z^*(\tau_{ij})<0)(1-\Delta_i)} \\ &\Pr(T_{\tau_{ij}} = t_i, Z^*_{\tau_{ij}} \geq 0; \boldsymbol{\theta})^{\mathbf{1}(Z^*(\tau_{ij})\geq 0)\Delta_i} \cdot \Pr(T_{\tau_{ij}} \geq t_i, Z^*_{\tau_{ij}} \geq 0; \boldsymbol{\theta})^{\mathbf{1}(Z^*(\tau_{ij})\geq 0)(1-\Delta_i)} \end{aligned} \tag{5}$$

where the likelihood contribution is given by one of the following for an individual at measurement time τ who:

- Has the event at time t and $Z(\tau) = 0$

$$\Pr(T_\tau = t, Z^*_\tau < 0; \boldsymbol{\theta}) = \frac{\partial}{\partial t} F_{T_\tau, Z^*_\tau}(t, 0; \boldsymbol{\theta}) = \Phi_2\left(\frac{q_2(0; \boldsymbol{\theta}_2) - \rho_\tau(\boldsymbol{\theta}_\rho) q_1(t; \boldsymbol{\theta}_1)}{\sqrt{1 - \rho_\tau(\boldsymbol{\theta}_\rho)^2}}\right) f_{T_\tau}(t; \boldsymbol{\theta}_1)$$

- Is alive or censored at time t and $Z(\tau) = 0$

$$\Pr(T_\tau \geq t, Z^*_\tau < 0; \boldsymbol{\theta}) = F_{Z^*_\tau}(0; \boldsymbol{\theta}_2) - F_{T_\tau, Z^*_\tau}(t, 0; \boldsymbol{\theta})$$

- Has the event at time t and $Z(\tau) = 1$

$$\Pr(T_\tau = t, Z^*_\tau \geq 0; \boldsymbol{\theta}) = \frac{\partial}{\partial t}[F_{T_\tau}(t; \boldsymbol{\theta}_1) - F_{T_\tau, Z^*_\tau}(t, 0; \boldsymbol{\theta})] = \Phi_2\left(-\frac{q_2(0; \boldsymbol{\theta}_2) - \rho_\tau(\boldsymbol{\theta}_\rho) q_1(t; \boldsymbol{\theta}_1)}{\sqrt{1 - \rho^2_\tau(\boldsymbol{\theta}_\rho)}}\right) f_{T_\tau}(t; \boldsymbol{\theta}_1)$$

- Is censored or still alive at time t and $Z(\tau) = 1$

$$\Pr(T_\tau \geq t, Z^*_\tau \geq 0; \boldsymbol{\theta}) = [1 - F_{Z^*_\tau}(0; \boldsymbol{\theta}_2)] - F_{T_\tau}(t; \boldsymbol{\theta}_1) + F_{T_\tau, Z^*_\tau}(t, 0; \boldsymbol{\theta})$$

where $q_1(t; \boldsymbol{\theta}_1) = \Phi^{-1}(F_{T_\tau}(t; \boldsymbol{\theta}_1))$ and $q_2(z; \boldsymbol{\theta}_2) = \Phi^{-1}(F_{Z^*_\tau}(z; \boldsymbol{\theta}_2))$.

Direct maximization of this pseudo-likelihood can be computationally intensive due to the potentially large number of parameters to be estimated and complexity of the chosen marginal models. Thus, we conduct estimation using the inference functions for margins (IFM) method.[18] First, the parameters $\widetilde{\boldsymbol{\theta}}_1$ and $\widetilde{\boldsymbol{\theta}}_2$ are estimated from their respective marginal models. Second, these estimates are held fixed in the pseudo-likelihood given by Eq.(5), $PL(\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2, \boldsymbol{\theta}_\rho)$, which is maximized over $\boldsymbol{\theta}_\rho$ to get $\widetilde{\boldsymbol{\theta}}_\rho$. The IFM estimate is then $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2, \widetilde{\boldsymbol{\theta}}_\rho)$, and the dynamic predictions of interest can be computed as $\Pr(T \geq \tau + s | T > \tau, Z(\tau) = z; \widetilde{\boldsymbol{\theta}})$ for $z = 0, 1$ from Eq.(3) and Eq.(4), respectively.

The standard errors for the marginal survival model parameters can be obtained using standard methods used for a Cox or parametric survival model.[19] The marginal marker model is estimated using repeated measurements from each individual, thus robust standard errors can be computed using a sandwich estimator.[20] Due to the use of a two-stage method for estimation, the analytic standard errors for the association parameters must account for the estimation variability of the marginal model parameters. Two-stage variance estimation for parametric and semiparametric copula models are presented in existing literature, but can result in complex expressions for flexible specifications of the marginal models.[18,21,22,23] Thus, a resampling scheme, such as jackknife[18] or bootstrapping[24], will be used to compute the standard errors of the association parameters.

3 | SIMULATION STUDY

We use a simulation study to assess the predictive performance of the proposed method and compare it to the existing dynamic prediction methods of joint modeling and landmarking. We focus on the situation where the binary marker starts at 0, and can change to 1, but changes from 1 to 0 are not possible.

3.1 | Performance comparison metrics

We compute the dynamic predictions at a sequence of prediction times τ for the probability of experiencing the event in the interval $(\tau, \tau + s]$, given by $\bar{p}_i(\tau, s) = 1 - p_i(\tau, s)$, where $p_i(\tau, s)$ is the dynamic prediction given in Eq.(1). We compare the dynamic predictions to the true conditional death probabilities, which are computed using the true parameter values to get the transition intensities that are then numerically integrated over the prediction window $[\tau, \tau + s)$.[14] At each prediction time τ, we compute the bias and variance of the dynamic predictions conditional on the marker value, i.e., $Z(\tau) = 0$ or $Z(\tau) = 1$. We evaluate calibration using the root mean squared prediction error (RMSE) between the true conditional death probabilities, $\bar{p}_{\text{True}}$, and the predictions obtained from each of the different models, $\bar{p}_{\text{Model}}$, conditional on the baseline covariates, given by

$$\text{RMSE}(\tau, s | \mathbf{X}) = \sqrt{\mathbb{E}\left[\left(\bar{p}_{\text{True},i}(\tau, s | \mathbf{X}) - \bar{p}_{\text{Model},i}(\tau, s | \mathbf{X})\right)^2\right]}$$

We evaluate the discrimination and overall performance of the dynamic predictions using dynamic versions of area under the curve (AUC) and Brier score (BS), which account for censoring. We denote these measures $\text{AUC}(\tau, s)$ and $\text{BS}(\tau, s)$, and use the following definitions presented in Blanche et al[25] for which inverse probability of censoring weight (IPCW) estimators are given,

$$\begin{aligned}
\text{AUC}(\tau, s) &= \Pr(\bar{p}_i(\tau, s) > \bar{p}_j(\tau, s) | D_i(\tau, s) = 1, D_j(\tau, s) = 0, T_i > \tau, T_j > \tau) \\
\text{BS}(\tau, s) &= \mathbb{E}\left[(D(\tau, s) - \bar{p}(\tau, s))^2 \, | T > \tau\right]
\end{aligned}$$

where $D_i(\tau, s) = \mathbb{1}_{(\tau < T_i \leq \tau + s)}$. Since BS depends on the cumulative incidence of death in the prediction window $(\tau, \tau + s]$, we use a standardized R^2-type measure that compares how well the predictions perform relative to predictions from a null model given by the Kaplan-Meier estimate, $\hat{\text{BS}}_0(\tau, s)$, which does not take into account subject-specific information. We denote this scaled measure $R^2(\tau, s) = 1 - \hat{\text{BS}}(\tau, s)/\hat{\text{BS}}_0(\tau, s)$. The measures of $\text{AUC}(\tau, s)$ and $\text{BS}(\tau, s)$ include all of the subjects who are alive at prediction time τ. To make comparisons between models, we compute the best-attainable AUC and R^2 using the predicted probabilities from the true models. We then examine the relative measures $\Delta\text{AUC} = \hat{\text{AUC}}_{\text{True}} - \hat{\text{AUC}}_{\text{Model}}$ and $\Delta R^2 = R^2_{\text{True}} - R^2_{\text{Model}}$ for each of the models, where values close to 0 indicate better performance.

For each scenario, we simulate 1000 subjects. A random sample of 500 subjects are selected for the training data set to which the models were fit. These models are then used to obtain dynamic predictions for the remaining 500 subjects who compose the validation data set. Performance metrics are computed for these predictions, and averaged across five hundred simulations.

3.2 | Simulation Setup

Using a similar scenario as in Suresh et al,[14] we simulate patients from an illness-death model, which is a joint model for a time-to-event outcome and a binary time-dependent covariate. Such data can arise when there is a intermediate event (e.g., illness) that can occur during patient follow-up prior to a terminal event (e.g., death). Thus, in our defined notation, T represents the time to the terminal event, and the marker process $Z(t)$ indicates whether the patient has experienced the intermediate event by time t. Defining the states as {0: Healthy, 1: Ill, 2: Dead}, the ages of illness onset and death without illness were generated from

$$\lambda_{jk}(t_i|\mathbf{X}_i) = \left(\frac{\rho_{jk}}{\kappa_{jk}}\right)\left(\frac{t_i}{\kappa_{jk}}\right)^{\rho_{jk}-1} \exp\{\boldsymbol{\alpha}'_{jk}\mathbf{X}_i\} \quad \text{for } j=0, k=1,2$$

For transition intensity from illness to death ($1 \rightarrow 2$), we generate data under two different models: (1) Markov, where the transition intensity depends only on current time, i.e., $\lambda_{12}(t|\mathbf{X})$, and (2) semi-Markov ("clock-reset"), where the transition depends on duration in the illness state i.e., $\lambda_{12}(t-V|\mathbf{X})$, where V is the known transition time. The change in the binary marker value from 0 to 1 corresponds to the healthy-to-ill transition and is determined by the hazard $\lambda_{01}(t)$. The other two transition intensities $\lambda_{02}(t)$ and $\lambda_{12}(t)$ represent the hazard function for death conditional on the marker value being 0 and 1, respectively.

We choose the transition intensity shape and scale parameters such that $\lambda_{12}(t) > \lambda_{02}(t) > \lambda_{01}(t)$ [$\rho_{jk} = 1.15$ for all $j \rightarrow k$, $\kappa_{01} = 15$; $\kappa_{02} = 12.5$; $\kappa_{12} = 10$], to achieve 25% of patients developing illness. We simulate a binary covariate X with prevalence 50%, that has a stronger effect on death in ill subjects, with $\alpha_{01} = 0.5$, $\alpha_{02} = 0.5$, $\alpha_{12} = 2$. We generate right-censoring from a Uniform(0,15) distribution to achieve a 50% censoring rate. We simulate marker measurement under two patterns of observation: (1) the marker process is continuously observed, and (2) the value of the marker is observed at random inspection times. Inter-inspection times are exponentially distributed with rate 0.5 and 1, to simulate both frequent and more sparsely collected marker measurements.

In addition to the basic scenario of a single baseline covariate, we also evaluated the performance of landmark models when the baseline covariate vector varies by transition. We generate data with two binary baseline covariates X_1 that has a stronger effect on death in ill subjects [$\alpha_{01,1} = \alpha_{02,1} = 0.5, \alpha_{12,1} = 2$] and X_2, which has no effect on death [$\alpha_{01,2} = 1, \alpha_{02,2} = \alpha_{12,2} = 0$]. We are interested in the dynamic prediction of failure at the prediction times $\tau = 0, 1, \ldots, 5$, for a prediction window of 3 years beyond the prediction time. A summary of the scenarios that were simulated under are given in Table 1.

TABLE 1 Summary of scenarios for simulation study.

Scenario	Model	Baseline covariates	Inter-inspection rate
1a	Markov	X	0.5
1b	Markov	X	1
1c	Markov	X	Continuously observed
2a	Semi-Markov	X	0.5
2b	Semi-Markov	X	1
2c	Semi-Markov	X	Continuously observed
3a	Markov	X_1, X_2	0.5
3b	Markov	X_1, X_2	1
3c	Markov	X_1, X_2	Continuously observed

3.2.1 | Models for Dynamic Prediction

In addition to the copula approach We fit Markov and semi-Markov joint models, and landmark models, as shown in Table 2. (MM) is a Markov illness-death model with Weibull transition intensities. (MSM) accounts for the effect of the observed transition time on the risk of death for those in the illness state. (MMCox) and (MSMCox) are their semiparametric counterparts. (SMM) is a parametric semi-Markov ("clock-reset") illness-death model, where the risk of transition to death after illness depends on the duration of time the individual has spent in the illness state. We also consider the flexible landmark models introduced in Suresh et al,[14] which can be fit to unbalanced longitudinal data. (LM3) is the extended super landmark model and allows for non-proportional hazards. (LM4) allows the covariate effects of illness status to be a function of both measurement time τ and residual time $t - \tau$. (LMInt3) and (LMInt4) extend these models to include an interaction term between illness status and the baseline covariates. These interaction models were found to have significantly improved performance over the regular landmarking models, especially when there were multiple baseline covariates with differential effects for the different transitions.[14] (LSM3) and (LSM4) are fit in the scenarios using the semi-Markov model for generating data, and account for the dependency of transition on the observed illness time by including it as a covariate.

To identify the functional forms of the copula models we examine goodness-of-fit statistics and perform model selection, as demonstrated in Supplementary Material B. We present the results from six flexible copula models. Failure time data is modeled either parametrically (W: Weibull) or semiparametrically (C: Cox) and the binary marker data is modeled using a probit regression. In (B*1), we model both the association and the mean of the continuous latent variable underlying the binary marker as a function of time and the baseline covariate. In (B*2), we increase the flexibility by including an interaction between the baseline covariate and time in the model for the mean of the latent variable. In (B*3), we consider an interaction between the baseline covariate and time in both the model for the marker and for the association. We also considered more flexible forms for the mean and association using splines and higher order terms, but found that the additional flexibility did not improve fit or performance. Since we simulate data from a joint model, the copula and landmark models in all of the scenarios are misspecified models. Prediction for all three methods computes the dynamic prediction probabilities conditional on the scalar marker value at the prediction time, using a last-observation-carried-forward imputation for inspection time scenarios. R code for estimating these models and the dynamic predictions is available at https://github.com/ksuresh17/binarymarker-copula-dyn-pred.

3.2.2 | Simulation Results

We present the simulation results comparing the three methods for dynamic prediction in Supplementary Material C. First, we simulate under a Markov assumption with a single baseline covariate, and in Figure 1 present the results from the inspection time measurement setting (Scenario 1a). As expected, the joint model from which the data were simulated (MM) has the best predictive performance. We find that the copula model has better RMSE for both values of the binary baseline covariate than the misspecified Cox model with semiparametric baseline hazards (MMCox) and the landmark models (LM3) and (LMInt3). We present the bias for $X = 1, Z(\tau) = 1$ (i.e., those in the illness group with baseline covariate $X = 1$ and who have transitioned to the illness state by prediction time τ), and find that as the prediction time increases the bias for the copula model worsens. At the later time points there are very few individuals in this group (3% at LM=5), demonstrating that the copula model does not fit the data well at later time points for groups that have sparse data at those times. The copula model has low variance and BS relative to the other models, and comparable AUC. As the inspection time increases (Scenario 1b, 1c), the performance of the landmark models with the interaction (LMInt*) and semiparametric Markov model (MMCox) improve to be on par with the copula model. The copula and other models consistently outperform the landmark models without the interaction term.

On average, in Scenario 1a the computation time for estimation for the joint models (MM) and (MMCox) took 10.4 and 0.02 seconds, respectively. The landmark models ranged from 1.62-1.93 seconds, with (LM3) and (LMInt3) having faster computation time than (LM4) and (LMInt4), but the models that included the interaction taking longer than those without. The copula models that included simple and interaction effects in the marker process (BC1), (BW1), (BC2), (BW2) took about 0.92 seconds, with the models that used a Weibull model for the failure time data taking slightly longer than those that used the Cox models. The copula models (BC3) and (BW3) that included an interaction in the association function took longer at 1.87 and 1.93 seconds, respectively. These relationships were consistent across the other simulation scenarios as well and are summarized in Supplementary Material Table C1.

The performance of the copula model fit with a semiparametric Cox model for the marginal survival time distribution (BC*) has higher RMSE than the parametric version (BW*) but performs similarly or slightly better for the other performance metrics. Comparing the copula models, the models that include additional flexibility in the model for the mean of the latent variable (B*2)

TABLE 2 Summary of models fit in the simulation study.

Class	Model	Label
Markov	$\lambda^{W}_{jk,0}(t)\exp\{\alpha_{jk}X\}$ for $j \to k$ transition	(MM)
Markov, V^*	$\lambda^{W}_{jk,0}(t)\exp\{\alpha_{jk}X + \gamma V^*\mathbf{1}(j=1,k=2)\}$	(MSM)
Semi-Markov	$\lambda^{W}_{jk,0}(t - V^*\mathbf{1}(j=1,k=2))\exp\{\alpha_{jk}X\}$	(SMM)
	$\lambda^{W}_{jk,0}(t)$ modeled as Weibull hazard	
Markov	$\lambda^{Cox}_{jk,0}(t)\exp\{\alpha_{jk}X\}$ for $j \to k$ transition	(MMCox)
Markov, V^*	$\lambda^{Cox}_{jk,0}(t)\exp\{\alpha_{jk}X + \gamma V^*\mathbf{1}(j=1,k=2)\}$	(MSMCox)
	$\lambda^{Cox}_{jk,0}(t)$ modeled nonparametrically	
Landmark Models	$h_0(t)\exp\{\theta(\tau) + \beta_0 Z(\tau) + \omega(t-\tau)Z(\tau) + \alpha X\}$	(LM3)
	$h_0(t)\exp\{\theta(\tau) + \beta_0 Z(\tau) + \omega(t-\tau)Z(\tau) + \alpha_1 X + \alpha_2 X Z(\tau)\}$	(LMInt3)
	$h_0(t)\exp\{\theta(\tau) + \beta_0 Z(\tau) + \omega(t-\tau)Z(\tau) + \gamma V^* Z(\tau) + \alpha X\}$	(LSM3)
	$h_0(t)\exp\{\theta(\tau) + \beta(\tau) Z(\tau) + \omega(t-\tau)Z(\tau) + \alpha X\}$	(LM4)
	$h_0(t)\exp\{\theta(\tau) + \beta(\tau) Z(\tau) + \omega(t-\tau)Z(\tau) + \alpha_1 X + \alpha_2 X Z(\tau)\}$	(LMInt4)
	$h_0(t)\exp\{\theta(\tau) + \beta(\tau) Z(\tau) + \omega(t-\tau)Z(\tau) + \gamma V^* Z(\tau) + \alpha X\}$	(LSM4)
Copula Models	C: Gaussian copula	
	$\mu_{Z^*} = \gamma_0 + \gamma_1\tau + \gamma_2 X$	
	$\eta_\tau = \xi_0 + \xi_1\tau + \xi_2 X$	
	$h(t) = h_0(t)\exp\{\nu X\}$; $h_0(t)$ modeled nonparametrically	(BC1)
	$h(t) = h_0(t)\exp\{\nu X\}$; $h_0(t)$ modeled as Weibull hazard	(BW1)
	C: Gaussian copula	
	$\mu_{Z^*} = \gamma_0 + \gamma_1\tau + \gamma_2 X + \gamma_3 X\tau$	
	$\eta_\tau = \xi_0 + \xi_1\tau + \xi_2 X$	
	$h(t) = h_0(t)\exp\{\nu X\}$; $h_0(t)$ modeled nonparametrically	(BC2)
	$h(t) = h_0(t)\exp\{\nu X\}$; $h_0(t)$ modeled as Weibull hazard	(BW2)
	C: Gaussian copula	
	$\mu_{Z^*} = \gamma_0 + \gamma_1\tau + \gamma_2 X$	
	$\eta_\tau = \xi_0 + \xi_1\tau + \xi_2 X + \xi_3 X\tau$	
	$h(t) = h_0(t)\exp\{\nu X\}$; $h_0(t)$ modeled nonparametrically	(BC3)
	$h(t) = h_0(t)\exp\{\nu X\}$; $h_0(t)$ modeled as Weibull hazard	(BW3)

V^*: observed illness time; X: baseline covariate vector; $Z(t)$: value of binary marker at time t (0: healthy; 1: ill); $\beta(\tau) = \beta_0 + \beta_1\tau + \beta_2\tau^2$; $\theta(\tau) = \theta_1\tau + \theta_1\tau^2$; $\omega(s) = \omega_1 s + \omega_2 s^2$

and in the association function (B*3) have almost identical performance to that of the simpler models (B*1). These relationships between the copula models holds across all of the simulation scenarios.

For the semi-Markov simulation setting, we compare the copula model with landmark models and joint models that condition on the observed transition to illness. We present the results for the inspection time measurement setting in Figure 2 (Scenario 2a). We find that the copula model has better performance than the landmark models and the semiparametric semi-Markov model (MSMCox). It has low variance and Brier score and has an AUC comparable with that of (SMM). As the inspection time increases (Scenario 2b, 2c), the performance of (MSMCox) improves, but the copula model still outperforms the landmark models across all the metrics.

Finally, we generate data under a Markov model with two baseline covariates that have differing effects for the different transitions. From Figure 3, in the setting with inspection time measurement (Scenario 3a) we see that the copula model has low variance and Brier score compared to the landmark models, and comparable RMSE to the landmark model with the interaction and the semiparametric Markov model. We present bias for the group $X_1 = 1, X_2 = 1, Z(\tau) = 1$, and find that for the copula model the bias increases with prediction time. Again, we find that this is associated with few people being in that group at later times, preventing the copula from estimating the marginal distributions well at those times.

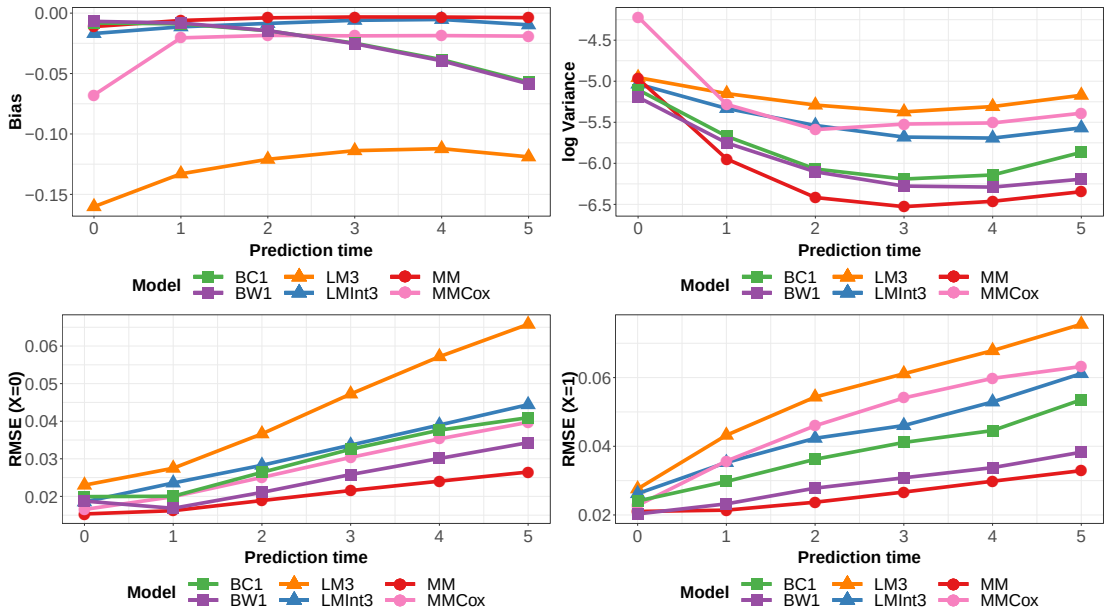

FIGURE 1 Simulation estimates for binary marker Scenario 1a (Markov illness-death model with one baseline covariate and inspection rate 0.5) for bias (upper-left) and variance (upper-right) for $Z(\tau) = 1, X = 1$, and RMSE for $X = 0$ (bottom-left) and $X = 1$ (bottom-right) for predicted probability $P(T \leq \tau + 3 | T > \tau, Z(\tau), X)$ from copula models (BC1), (BW1), joint models (MM), (MMCox) and landmark models (LM3), (LMInt3).

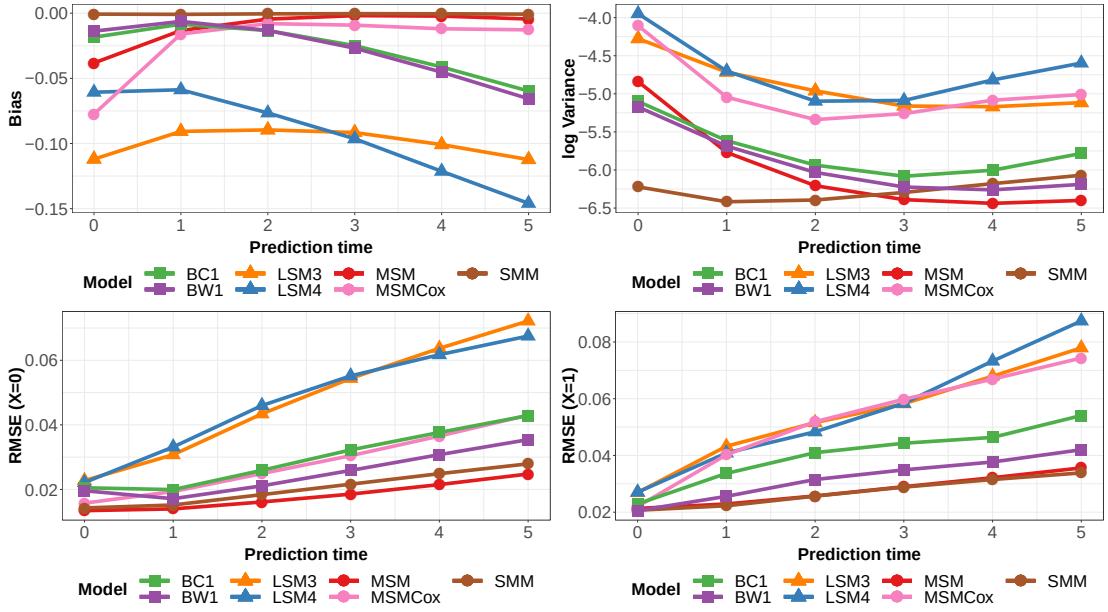

FIGURE 2 Simulation estimates for binary marker Scenario 2a (semi-Markov illness-death model with one baseline covariate and inspection rate 0.5) for bias (upper-left) and variance (upper-right) for $Z(\tau) = 1, X = 1$, and RMSE for $X = 0$ (bottom-left) and $X = 1$ (bottom-right) for predicted probability $P(T \leq \tau + 3 | T > \tau, Z(\tau), X)$ from copula models (BC1), (BW1), joint models (MSM), (MSMCox), (SMM), and landmark models (LSM3), (LSM4).

Overall, the copula model has good predictive performance across all the metrics, performing better than landmark models and misspecified Markov models with less frequent inspection times, and on par with other models with a continuously observed binary marker. The copula model consistently outperforms the landmark model without the interaction term, indicating that it has

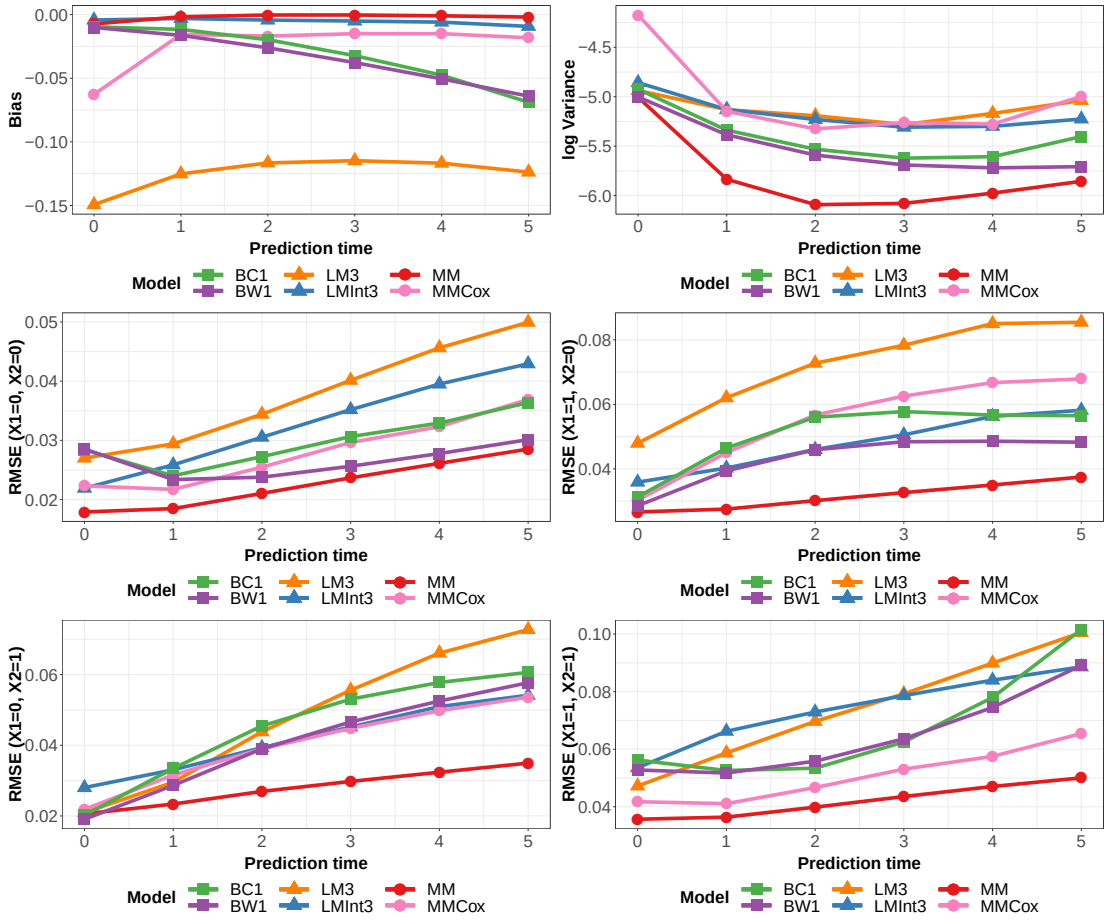

FIGURE 3 Simulation estimates for binary marker Scenario 3a (Markov illness-death model with two baseline covariates and inspection rate 0.5) for bias and variance for $Z(\tau) = 1, X_1 = 1, X_2 = 1$, and RMSE for predicted probability $P(T \leq \tau + 3 | T > \tau, Z(\tau), \mathbf{X})$ from copula models (BC1), (BW1), joint models (MM), (MMCox) and landmark models (LM3), (LMInt3).

better predictive performance than the standard landmark models that do not include additional flexibility. The bias for the copula model can be high for groups at times where there is little data observed; however, from RMSE we see that overall performance of the copula model is better or comparable to the flexible landmark and misspecified Markov models. In comparing the copula models, as in the continuous marker situation, we find that changes in the association structure result in similar predictive performance.[8] This suggests that with well-chosen models for the marginal latent marker and failure time distributions, flexible association functions can be specified.

4 | APPLICATION: PROSTATE CANCER STUDY

Returning to the prostate cancer study in Suresh et al,[14] we demonstrate and assess the use of the copula model for obtaining dynamic predictions using a binary marker. The data consists of 745 patients with clinically localized prostate cancer who were treated with radiation therapy. Patients were followed from start of treatment (baseline) and monitored for the occurrence of metastatic clinical failure (CF), treated as a time-dependent binary covariate. The aim is to use the intermediate CF information to predict a patient's future risk of death. The median follow-up time was 9 years, and 52 patients experienced CF during the study. Out of 188 total deaths, 154 patients died before and 34 died after experiencing clinical failure. The pretreatment prognostic factors measured at baseline are continuous age (median 69; IQR 63-74), log(PSA + 1) (PSA ng/ml; median 8; IQR 5-12), and Gleason score with 7="3+4" and 7.5="4+3" (median 7; IQR 6-7.5), and categorical prostate cancer stage (T1: 57%, T2-T3:

43%), and number of comorbidities (0: 55%, 1-2: 37%, ≥3: 8%). We are interested predicting the probability of death within 5 years at prediction times $\tau = 0, 1, \ldots, 8$ years following start of treatment.

After performing model selection and assessing goodness-of-fit, we fit the following Gaussian copula model: $h(t) = h_0(t)\exp\{\boldsymbol{\nu}\mathbf{X}\}$, $\mu_{Z^*} = \gamma_0 + \boldsymbol{\gamma}_1\mathbf{X} + \sum_{k=1}^{3}\gamma_{2k}B_k(\tau)$, $\eta_\tau = \xi_0 + \boldsymbol{\xi}_1\mathbf{X} + \sum_{i=1}^{3}B_k(\tau, \boldsymbol{\xi}_2)$, where $\mathbf{X}$ is a vector of the baseline covariates, B_k is a B-spline for a natural cubic spline with boundary knots at 0 and 10 years. We consider failure time models where $h_0(t)$ is modeled nonparametrically (CopCox) and parametrically with a Weibull baseline hazard (CopWeib), and model the binary marker data using a probit regression. We evaluate the fit of the Cox model to the failure time data, and find that there is no violation of the proportional hazards assumption for any of the baseline covariates. We assess the fit of the probit model to the binary marker and identify that no covariate transformation is required. The model for the association parameter function was chosen to be a flexible function of measurement time and baseline covariates. Details for assessing goodness of fit are given in Supplementary Material D.

The parameter estimates for the components of the copula model are given in Table 3. Robust standard errors were computed for the marginal marker model coefficient estimates, and standard errors for the association parameters were computed using bootstrapping. Additionally, we fit joint and landmark models explored in our simulation study, and present results from the parametric and semiparametric joint models (MM) and (MMCox), and the extended super landmark models (LM4, LMInt4). The parameter estimates for these models are given in Supplementary Material D.

For the marginal model for time to death, increased age, PSA, Gleason score, and number of comorbidities are significantly associated with increased risk of death. From the marginal model for the binary marker data, increased age, Gleason score, and Stage T2-T3 were associated with increased probability of developing CF. These relationships were also observed in the joint models. Unlike the copula model, the landmark models are not able to evaluate the effect of the baseline covariates on the risk of CF. The bootstrapped association parameter standard errors are large due to the incorporation of the estimation uncertainty of the first-stage parameters. But negative association parameter estimates suggest that increasing Gleason score and Stage T2-T3 result in more negative association between the latent variable underlying CF and time to death, indicating that patients with those characteristics have high negative association between CF and death (i.e., decreased time to death). Similarly, the positive coefficient for having 1-2 comorbidities compared to 0 comorbidities indicates positive association between CF and time to death, and thus decreased risk of death. This relationship was also demonstrated in the landmark models with interactions.

In Figure 4, we present the predicted probabilities for two individuals in the data set from the copula, landmark, and joint models. Individual A is at increased risk of death due to risk factors (older, increased PSA, high Gleason score), but does not experience clinical failure before death. Individual B is a lower risk patient, but has some baseline characteristics (increased PSA, high Gleason score) that indicate increased probability of CF, and that greatly increase his risk of death after experiencing clinical failure. In the probability plots, the predictions from the copula models are very similar to the joint models, (MM) and (MMCox), and the landmark model with the interaction (LMInt4). Unlike the landmark model without the interaction (LM4), the copula model is able to take into account the differential effects of the baseline covariates on the different transitions, which is demonstrated by the large increase in predicted probability of death after CF for Individual B. There is no difference in the predicted probabilities for (CopCox) and (CopWeib) for Individual A, but we see that the predictions from (CopWeib) are lower than (CopCox) in Individual B after they experience CF. In Figure 5, we present the association functions for the two individuals. As prediction time increases the association between time to death and CF is negative but is increasing and approaches zero, thus indicating that as time from treatment increases the predicted probability of death relies less on an individual's CF status. This is also demonstrated in the effect of the interaction between CF and measurement time in the landmark models where as the prediction time increases the effect of CF on the risk of death decreases.

5 | DISCUSSION

Dynamic prediction methods that incorporate the effect of a patient's changing longitudinal information into their survival prediction are necessary for making informed, and personalized treatment decisions. Existing methods for dynamic prediction are often focused on incorporating continuous marker information; however, often binary indicators that identify the occurrence of an intermediate event can be collected during follow-up. We propose a Gaussian copula approach for dynamic prediction of survival that incorporates binary time-dependent information collected during follow-up.

The Gaussian copula approach for dynamic prediction has been shown in previous work to have good predictive performance in the continuous marker setting.[8] By separately modeling the marginal marker and survival data and their association, it has

TABLE 3 Coefficient estimates and standard errors for copula model applied to prostate cancer data with binary marker of metastatic clinical failure.

		CopCox		CopWeib	
	Covariate	Coef.	SE	Coef.	SE
ν	Age	0.073	0.012	0.071	0.012
	log(PSA+1)	0.263	0.110	0.261	0.110
	Gleason Score	0.311	0.084	0.283	0.082
	Stage T2-T3	0.043	0.158	0.114	0.156
	Comorbidities 1-2	0.472	0.163	0.468	0.162
	Comorbidities ≥ 3	1.228	0.217	1.204	0.216
γ	Intercept	-6.152	1.074	Same parameter estimates and SEs as CopCox	
	Age	0.002	0.012		
	log(PSA+1)	0.267	0.075		
	Gleason Score	0.220	0.109		
	Stage T2-T3	0.245	0.175		
	Comorbidities 1-2	0.096	0.188		
	Comorbidities ≥ 3	-0.120	0.280		
	B_1	2.523	0.553		
	B_2	1.416	0.371		
	B_3	1.713	0.323		
ξ	Intercept	-0.498	2.332	-0.283	2.069
	Age	0.005	0.016	0.007	0.015
	log(PSA+1)	0.024	0.228	-0.020	0.192
	Gleason Score	-0.151	0.191	-0.147	0.171
	Stage T2-T3	-0.314	0.396	-0.285	0.384
	Comorbidities 1-2	0.230	0.312	0.225	0.284
	Comorbidities ≥ 3	-0.117	0.402	-0.006	0.311
	B_1	1.789	2.219	1.105	1.871
	B_2	0.050	0.888	-0.079	0.765
	B_3	1.207	1.266	0.825	1.059

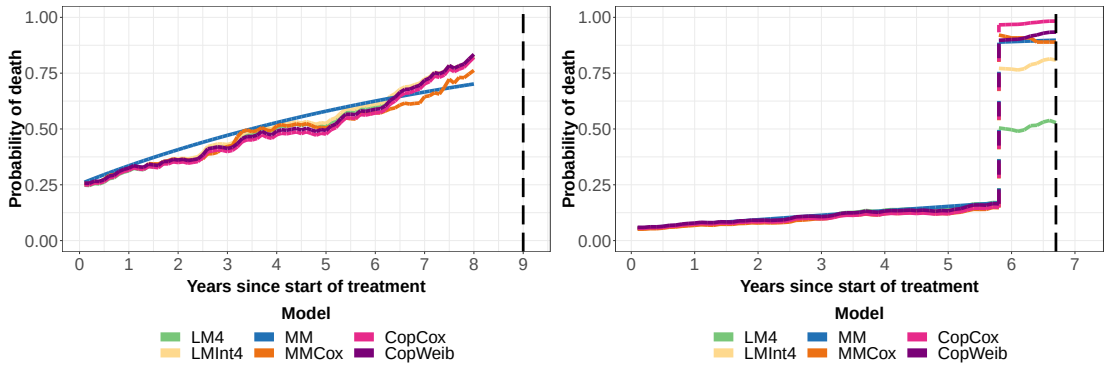

FIGURE 4 Predicted probability of death within 5 years, $P(T \leq \tau+5|T > \tau, Z(\tau), \mathbf{X})$ for two individuals in the prostate cancer data set for landmark, joint, and copula models. Individual A (left) is 75 years old at baseline, with PSA 29.9 ng/mL, Gleason score 9, T2 Stage, 2 comorbidities, and does not experience clinical failure but dies 9 years from baseline. Individual B (right) is 67 years old at baseline, with PSA 12.6 ng/mL, Gleason score 8, T1 Stage, zero comorbidities, and experiences clinical failure 5.8 years after start of treatment before dying at time 6.7 years from baseline. Black dashed line indicates time of death.

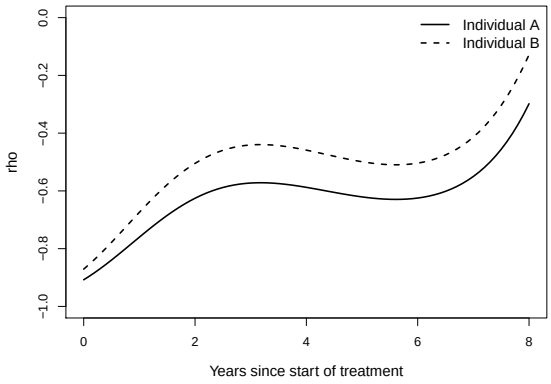

FIGURE 5 Association functions from (CopCox) for Individual A (solid) and Individual B (dashed) from the prostate cancer data set.

the advantage of allowing us to assess goodness-of-fit and perform variable selection to minimize bias at the marginal model stage. Unlike landmarking, it does not require fixing the prediction horizon and the prediction times of interest for estimation. In comparison to more complex joint models, estimation can be performed using standard software, and the dynamic predictions of interest are easily derived.

Since the Gaussian copula is only applicable for modeling the joint distribution of two continuous outcomes, using a latent variable formulation we extend its use for the binary marker setting. We demonstrate that its predictive performance is on par with those of joint modeling and landmarking under various scenarios, and show its use for obtaining dynamic predictions in a data application. This approach provides us with an alternative method for dynamic prediction when incorporating a time-dependent binary covariate, with advantages over the existing methods of landmarking and joint modeling.

A limitation of the Gaussian copula approach is that since it models the joint distribution of the marker and survival conditional on surviving to the prediction time, it relies on the availability of data at those prediction times. In the binary marker simulations, we demonstrate that as the number of people in a particular group decreases over time (due to death or censoring), the bias of the predictions for that group increases. In addition, the large standard errors in the association function, resulting from the two-stage estimation approach, make it difficult to perform variable selection for identifying the optimal association function specification. With this approach we have to specify a functional form for the marker, the survival, and their association based on covariates. However, from the simulation study we find that the predictive performance of the copula is similar when comparing flexible functions for these components.

Using a copula framework provides the potential for several extensions to more complicated data structures. In this paper, we mainly consider a single binary time-dependent variable that can transition from 0 to 1 during a patient's follow-up. The use of the copula to describe the distribution of the latent marker value over time suggests an easy extension to more complex data structures, such as when the patient's binary marker can transition from 0 to 1 and back multiple times. We can then also include as a covariate the number of reversals a patient has experienced by a particular prediction time in the models for the conditional marker and/or residual time distributions to account for increased risk of the binary marker and/or death. Additional summary variables of the binary marker up to the prediction time, such as time spent in the illness state, can also be similarly included in the different components of the model.

Dynamic predictions are usually implemented in longitudinal studies where dropout is a common complication. This dropout may be random or it may be associated with the longitudinal variable (making it missing at random, MAR) or there may be a form of dependent censoring in which the dropout is related to the event (making it missing not at random, MNAR). If the data set does have this feature then an interesting question is how well the three approaches will behave under these type of dropout scenarios. We speculate that all three approaches would work under completely random dropout. In previous work,[8] we have demonstrated that a copula approach for dynamic prediction has similar performance to joint modeling when missingness of the longitudinal marker is dependent on observed variables. Under MAR we would expect the joint modeling approach to continue to work well because it is based on a likelihood from a unified model. Whether and by how much the performance of the copula

and landmarking approaches will deteriorate under MAR will likely depend on the exact scenario. All approaches are likely to behave less well if the dropout is MNAR.

The copula formulation also allows us to extend from a bivariate copula to a multivariate copula to accommodate multiple longitudinal markers. By adapting the Gaussian copula approach for dynamic prediction to a binary marker setting, we can use a multivariate copula to incorporate both the effect of binary and continuous markers into updating a patient's prediction. We can model the various markers using appropriate marginal distributions based on their specific data types, and separately describe their association with the failure time using the copula. This approach would replace the association function with an association matrix, which would also allow us to account for the correlation between the multiple longitudinal markers. Although, care should be taken to propose parsimonious models for the marginals and the association functions to avoid exponentially increasing the number of parameters to estimate.

With this work we have demonstrated that an approximate approach that models only a component of the joint distribution of the marker and survival process can incorporate additional flexibility to achieve good predictive performance. In complex settings, joint models may be difficult to specify and estimate, and their predictive performance is sensitive to misspecification. Future work will explore developing and extending approximate approaches for dynamic prediction with complicated data structures, such as interval censoring and multiple longitudinal markers.

ACKNOWLEDGEMENTS

This research was partially supported by National Institutes of Health grants CA199338 and CA129102.

References

1. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* 2011; 67(3): 819–829.
2. Rizopoulos D, Murawska M, Andrinopoulou ER, Molenberghs G, Takkenberg JJ, Lesaffre E. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. arXiv:1306.6479 [stat.AP]; 2013.
3. Rizopoulos D, Molenberghs G, Lesaffre EM. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal* 2017; 59(6): 1261–1276.
4. Houwelingen vH, Putter H. *Dynamic prediction in clinical survival analysis*. CRC Press . 2011.
5. Putter H. dynpred: Companion package to "Dynamic prediction in clinical survival analysis.". *R package version 0.1* 2015; 2.
6. Jewell NP, Nielsen JP. A framework for consistent prediction rules based on markers. *Biometrika* 1993; 80(1): 153–164.
7. Ferrer L, Putter H, Proust-Lima C. Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment. *Statistical methods in medical research* 2019; 28(12): 3649–3666.
8. Suresh K, Taylor JM, Tsodikov A. A Gaussian copula approach for dynamic prediction of survival with a longitudinal biomarker. *Biostatistics* 2019.
9. Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical models based on counting processes*. Springer Science & Business Media . 2012.
10. Hougaard P. Multi-state models: a review. *Lifetime data analysis* 1999; 5(3): 239–264.
11. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 1978: 141–150.
12. Fiocco M, Putter H, Houwelingen vHC. Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statistics in Medicine* 2008; 27(21): 4340–4358.

13. Van Houwelingen HC, Putter H. Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime data analysis* 2008; 14(4): 447.

14. Suresh K, Taylor JM, Spratt DE, Daignault S, Tsodikov A. Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal* 2017; 59(6): 1277–1300.

15. Song PXK, Li M, Yuan Y. Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* 2009; 65(1): 60–68.

16. Leon dAR, Wu B. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine* 2011; 30(2): 175–185.

17. Sklar A. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* 1959; 8: 229-231.

18. Joe H, Xu JJ. The estimation method of inference functions for margins for multivariate models. *Technical Report* 1996.

19. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* 1982: 1100–1120.

20. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986: 121–130.

21. Song XK. *Correlated data analysis: Modeling, analytics, and applications*. Springer Science & Business Media . 2007.

22. Spiekerman CF, Lin D. Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association* 1998; 93(443): 1164–1175.

23. Prenen L, Braekers R, Duchateau L. Extending the Archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2017; 79(2): 483–505.

24. Efron B, Tibshirani RJ. *An introduction to the bootstrap*. CRC press . 1994.

25. Blanche P, Proust-Lima C, Loubère L, Berr C, Dartigues JF, Jacqmin-Gadda H. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* 2015; 71(1): 102–113.