

Dynamic Prediction of Renal Failure Using Longitudinal Biomarkers in a Cohort Study of Chronic Kidney Disease

Liang Li · Sheng Luo · Bo Hu · Tom Greene

Received: October 16, 2015 / Accepted:

Abstract In longitudinal studies, prognostic biomarkers are often measured longitudinally. It is of both scientific and clinical interest to predict the risk of clinical events, such as disease progression or death, using these longitudinal biomarkers as well as other time-dependent and time-independent information about the patient. The prediction is dynamic in the sense that it can be made at any time during the follow-up, adapting to the changing at-risk population and incorporating the most recent longitudinal data. One approach is to build a joint model of longitudinal predictor variables and time to the clinical event, and draw predictions from the posterior distribution of the time to event conditional on longitudinal history. Another approach is to use the landmark model, which is a system of prediction models that evolve with the follow-up time. We review the pros and cons of the two approaches, and present a general analytical framework using the landmark approach. The proposed framework allows the measurement times of longitudinal data to be irregularly spaced and differ between subjects. We propose a unified kernel weighting approach for estimating the model parameters, calculating predicted probabilities, and evaluating prediction accuracy through double time-dependent Receiver Op-

This research is supported by NIH grants P30CA016672 and 5R01DK090046. The authors thank the Guest Editor and two anonymous reviewers for insightful comments that greatly improved this paper.

Liang Li

Department of Biostatistics, University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, USA. E-mail: LLi15@mdanderson.org

Sheng Luo

Department of Biostatistics, University of Texas School of Public Health, Houston, TX, USA

Bo Hu

Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA

Tom Greene

Department of Population Health Sciences, University of Utah, Salt Lake City, UT, USA

erating Characteristics (ROC) curves. We illustrate the proposed analytical framework using the African American Study of Kidney Disease and Hypertension (AASK) to develop a landmark model for dynamic prediction of end stage renal diseases or death among patients with chronic kidney disease.

Keywords Biomarker · Joint modeling of longitudinal and survival data · Landmark analysis · Prediction · Time-dependent ROC · Varying coefficient model

1 Introduction

Chronic kidney disease (CKD) is characterized by progressive loss of kidney function and is estimated to affect 13.6% of the adults in the United States [41]. Kidney function is usually measured by the glomerular filtration rate (GFR), which estimates how much blood passes through the glomeruli each minute. Glomeruli are the tiny filters in the kidneys that filter waste from the blood. According to National Kidney Foundation guidelines [25], the normal GFR level for adults is greater than $90\text{mL}/\text{min}/1.73\text{m}^2$, and lower GFR indicates progressively more severe stages of CKD, which may finally lead to end stage renal disease (ESRD, including dialysis, kidney transplantation and other renal replacement therapies) or death. Our recent research demonstrates considerable heterogeneity in the longitudinal progression trajectories of GFR among patients with CKD [18]. While the GFR of CKD patients declines in general, the individual trajectories vary, with notable periods of stabilization, accelerated or decelerated decline or increase. The diversity of the GFR trajectory patterns makes it difficult to predict the future risk of ESRD or death. However, such prediction is important for both physicians and patients to properly manage the treatment of disease. For example, The KDIGO guidelines recommended the use of risk prediction models to help determine the appropriate time to prepare for renal replacement therapy [16].

Many risk prediction models have been proposed for CKD [8, 24, 37]. From a statistical perspective, these are all static prediction models in the sense that the predictors were measured at a fixed time, often the baseline assessment of the cohort being analyzed, and a regression model, *e.g.*, Cox model, is used to relate the predictors to the time from the fixed time point to the subsequent onset of the clinical event. This approach has several limitations when applied to the aforementioned prediction problem in CKD research. First, since CKD is a chronic condition, the follow-up can be long and the adverse event of interest and baseline may be many years apart, significantly attenuating any association between the baseline predictors and the outcome. This issue becomes particularly relevant with nonlinear progression patterns of GFR. Second, a prediction model developed by regressing the adverse event outcome on baseline predictors is applicable only to prediction made on a new patient at baseline; it does not apply, for example, when the prediction needs to be made on a new patient who lived two years after the baseline without the adverse event. The at-risk patient populations (patients who have not had

the adverse event at the time of prediction) may be different between baseline and two years later, with some high risk subjects dropping out due to the adverse events. The strength of association between predictors and outcome could also vary over time. Third, the predictors of the static model include only the baseline variables, but not the longitudinal history, which may be unknown or undefined at baseline. However, for prediction in the longitudinal context, it is often desirable to incorporate in the model the most recent clinical history, including the rate of change and volatility of biomarkers, medication history, hospitalization episodes, quality of life improvements, etc. These quantities are time-varying and their distributions may vary with the at-risk population. These issues are not addressed in the static prediction model.

The *dynamic prediction model* [29, 44] is a desirable choice for the CKD research problem above. It predicts the future risk of adverse events at any time, incorporating all the time-dependent predictive information about the patient up to the time of prediction. The model parameters are allowed to vary with the ever-changing at-risk population and possible time-varying association between predictors and outcome. There are mainly two distinct lines of approaches in the literature for dynamic prediction: joint modeling [29, 30, 38] and landmarking [42, 44, 48].

The typical joint modeling (JM) approach generates prediction from a Cox model with time-dependent covariates. Of note, usually such model cannot be used for prediction because the hazard at a future time depends on the covariate trajectory at that time, which is unknown at the end of prediction. The JM approach avoids this problem by modeling the longitudinal trajectory of the time-dependent covariates, typically through a subject-specific parametric function. In the context of CKD studies, the risk factors of ESRD or death include, among others, the GFR, proteinuria, albuminuria, medication history, recent episodes of acute kidney injury (AKI), hospitalization, blood pressure and diabetes control, etc. All of them are time-varying. It is nearly impossible to correctly specify a joint statistical model for all of their trajectories at the patient level. Second, even if a small number of time-varying risk factors are involved, the computation in fitting a joint model may be prohibitive or unstable [15, 30, 32, 39, 40, 46].

In this paper, we propose innovative development along the lines of the landmark approach, originally studied by Zheng and Heagerty [48] and Van Houweilingen [42]. The term “landmark” came from the landmark analysis for survival data [1, 7, 9]. A comprehensive review is available [44]. The idea of landmarking has also been used for predicting long-term survival using short-term event time [26] and for estimating the effect of a time-dependent covariate on treatment-free survival when the treatment initiation and survival outcomes are correlated [12]. For predicting right censored events using longitudinal data, this approach uses survival regression models relating the predictor variables measured at or prior to the time of prediction to the time gap from the prediction time to the outcome event of interest. It is computationally much simpler than joint modeling, and can handle large number of time-varying predictors without excessive computation difficulties. Therefore,

the scope of applications of the landmark method is much wider than that of joint modeling. While this paper was motivated by problems from CKD research, our approach is applicable to typical longitudinal cohort studies in which longitudinal data are collected from recurring clinical visits until the occurrence of the adverse event of interest or the time of random right censoring, whichever comes first. We allow the clinical visit times to be irregularly spaced and differ by subjects. The proposed methodology differs in important aspects from both the “super Cox model” approach [42, 44] and the “partly conditional survival model” [48], the two published lines of research in landmark prediction. Within our proposed framework, the construction of the landmark data set, the working model, and the predicted probability are all unified by a kernel weighting approach (Section 2). In Section 3, we further extend kernel weighting to model evaluation by proposing a double time-dependent Receiver Operating Characteristics (ROC) curve and associated estimation procedure. The proposed methodology is illustrated in Section 4 with data from a CKD cohort study. Discussions are included in Section 5.

2 Dynamic Prediction Model and Estimation

2.1 The notation and landmark data set

Let $i = 1, 2, \dots, n$ index the n subjects in the training data set from which the dynamic prediction model is to be developed. For the i -th subject, the time of the event of interest is denoted by T_i and the censoring time is denoted by C_i . We observe $Y_i = \min(T_i, C_i)$ and the censoring indicator $\delta_i = 1\{T_i \leq C_i\}$. Denote by \mathbf{Z}_{ij} , or equivalently $\mathbf{Z}_i(t_{ij})$, a vector of variables measured on subject i at time t_{ij} , $j = 1, 2, \dots, n_i$ and $t_{i1} < t_{i2} < \dots < t_{in_i} < Y_i$. The notation \mathbf{Z}_{ij} includes both the time-dependent and time-independent variables. For a time-independent variable, $Z_{ij} = Z_{ij'}$ for any $j \neq j'$. We assume that (1) the data from different subjects are independent and identically distributed; (2) for a given subject, the measurement times $\{t_{ij}\}$ are independent of both the longitudinal data $\{\mathbf{Z}_{ij}\}$ and survival time T_i ; (3) C_i is independent of T_i , $\{t_{ij}\}$, and $\{\mathbf{Z}_{ij}\}$. The measurement times from different subjects do not need to follow the same schedule; they can either be on a fixed grid or irregularly spaced on a continuous time scale.

Suppose at any time s , we want to develop a dynamic prediction model to calculate the probability of the event of interest in the next time window $(s, s + \tau_1]$ using the subject’s data collected in the history window $[s - \tau_2, s]$:

$$P(T_i \in (s, s + \tau_1] \mid Y_i > s, \mathbf{Z}_{ij} \text{ with } t_{ij} \in [s - \tau_2, s], j = 1, 2, \dots, n_i). \quad (1)$$

Van Houwelingen (2007) called τ_1 the *prediction horizon* [42]. We can also call the interval $(s, s + \tau_1]$ a *prediction window*. It quantifies how far into the future we want to predict. Its choice depends on the scientific question under study and needs to be specified prior to model development. If τ_1 is too small, then the prediction is for the immediate future, which may not

be of great interest for chronic disease prediction; if τ_1 is too large, then the prediction is for the distant future, which may not be accurate or justifiable for clinical events that respond quickly to a change in biomarkers and other longitudinal information. Sometimes, the research interest is to estimate the entire conditional distribution of the event given the history up to the time of prediction [48]. Conceptually, we can cast that problem into the framework of (1) by viewing τ_1 as any time point in $(0, \infty)$.

To make a prediction at any t_{ij} , when new longitudinal data just become available, we define $T_{ij} = T_i - t_{ij}$ and $C_{ij} = C_i - t_{ij}$ to be the residual times to event and censoring starting from t_{ij} . For prediction up to the horizon τ_1 , we artificially censor the residual survival time T_{ij} at τ_1 . Hence, at t_{ij} the derived time to event up to the horizon is $Y_{ij} = \min(T_{ij}, C_{ij}, \tau_1)$ and the derived censoring indicator is $\delta_{ij} = 1\{T_{ij} \leq C_{ij} \text{ and } T_{ij} \leq \tau_1\}$. T_{ij} and C_{ij} are independent given t_{ij} . Despite causing an increase in censoring rate, artificially censoring the residual survival at τ_1 remains desirable because it reduces the chance of model misspecification. For example, with a relatively short prediction horizon, one may not need to use a Cox model with an unspecified baseline hazard; a parametric survival model may provide an adequate fit for the time period within the horizon. This may increase the efficiency of estimation. Even if a Cox model is used, the concern over non-proportional hazards is greatly reduced with artificial censoring, making it plausible to avoid complicated time-varying coefficients modeling [48].

We call the interval $[s - \tau_2, s]$ a *history window*, which quantifies the amount of the past information that we believe is relevant to the prediction of the future. τ_2 can be infinity, in which case we use the entire longitudinal history of the patient for prediction. In problems where the most predictive data are in the immediate past and the data from the remote past are either redundant or out-dated, a small τ_2 may be used. We may also allow different τ_2 for different predictor variables. Both τ_1 and τ_2 can be made dependent on s , the prediction time. When s is close to baseline, i.e., $s < \tau_2$, we may need to set $\tau_2(s) = \min\{s, \tau_2\}$. When s approaches the end of the follow-up in the training data, τ_1 may be reduced properly to ensure identifiability.

Since the longitudinal measurement times may be irregularly spaced, different subjects may have varying number of longitudinal measurements falling in the history window. Therefore, it is often convenient to summarize the longitudinal data within the history window to create numeric predictors for the prediction model. Let \mathbf{X}_{ij} be the vector of derived predictor variables at the prediction time t_{ij} . \mathbf{X}_{ij} is a mapping from $\{\mathbf{Z}_{ij'} | t_{ij'} \in [t_{ij} - \tau_2, t_{ij}], j' = 1, 2, \dots, n_i\}$. Examples of possible mappings include: (1) the current value of the time-dependent variable, i.e., $\mathbf{X}_{ij} = \mathbf{Z}_{ij}$, (2) the average in the history window, i.e., $\mathbf{X}_{ij} = \sum_{j'} \mathbf{Z}_{ij'} 1\{t_{ij'} \in [t_{ij} - \tau_2, t_{ij}]\} / \sum_{j'} 1\{t_{ij'} \in [t_{ij} - \tau_2, t_{ij}]\}$, (3) a binary indicator of whether any of the measurements in the history window falls below a scientifically relevant threshold [29], (4) the slope, i.e., rate of change, of a time-dependent variable in the history window. Since some time-dependent variables may not be considered as predictive and used in the model, and some other variables may contribute to multiple mappings (e.g.,

we might simultaneously use the average, slope and volatility of a longitudinal biomarker in the history window), the dimension of \mathbf{X}_{ij} is not necessarily the same as \mathbf{Z}_{ij} . Since t_{ij} 's are irregularly spaced, there could be subjects who do not have any longitudinal data in the history window, leading to missing data in the corresponding \mathbf{X}_{ij} . Excluding these missing records will not cause bias in the model estimation procedure below, due to the assumption (2) at the beginning of this subsection.

We have defined the derived time to event variables (Y_{ij}, δ_{ij}) and derived longitudinal predictors \mathbf{X}_{ij} at the measurement times t_{ij} . We call the data set of derived outcome and predictor variables $\{Y_{ij}, \delta_{ij}, t_{ij}, \mathbf{X}_{ij} \mid i = 1, 2, \dots, n; j = 1, 2, \dots, n_i\}$ a *landmark data set*. It consists of $\sum_{i=1}^n n_i$ data pieces of $(Y_{ij}, \delta_{ij}, t_{ij}, \mathbf{X}_{ij}^T)^T$. In the methodological framework of this paper, each t_{ij} is both a *landmark time*, as defined below, and a prediction time, though the prediction can also be made at other times. It can also be viewed as a new baseline time for each data piece, as the derived time to event Y_{ij} starts from zero at these time points. This construction of the landmark data set is different from the ‘‘super Cox model’’ approach [42, 44]. In that approach, a series of landmark times $\eta_1, \eta_2, \dots, \eta_K$ are pre-specified. For each landmark time, the predictor variables are defined from longitudinal data measured before or at that time, and the outcome is the time gap from the landmark to the event of interest. A ‘‘super Cox model’’ is fit to the K landmark data sets stacked on top of each other. One limitation of the super Cox model approach is that it involves subjective choice of K and the landmark times $\{\eta_1, \eta_2, \dots, \eta_K\}$, and there is little guideline on the optimal way to do this. Another limitation is that with irregularly spaced measurement times, η_k may not coincide with any t_{ij} 's. Hence it is unclear how $\mathbf{X}_i(\eta_k)$ should be defined if we are interested in using the current (at the time of prediction) value of the biomarker to make prediction. The approach in this paper avoids the first limitation by using all the measurement times t_{ij} as landmark times. It avoids the second limitation by defining \mathbf{X}_{ij} , Y_{ij} and δ_{ij} in the training data only at t_{ij} , synchronizing all history windows and prediction windows at times when new longitudinal measurements become available. This brings more rigor to the way predictors are defined and prediction probabilities are calculated (Section 2.2).

2.2 The landmark model and predicted probabilities

We fit the landmark data set $\{Y_{ij}, \delta_{ij}, t_{ij}, \mathbf{X}_{ij} \mid i = 1, 2, \dots, n; j = 1, 2, \dots, n_i\}$ to the following working model: at any measurement time t_{ij} and given the predictor variables in \mathbf{X}_{ij} , T_{ij} arises from a distribution with the hazard function

$$\begin{aligned} h_i(t \mid t_{ij}, \mathbf{X}_{ij}) &= \lambda_0(t, t_{ij}) \exp \{ \beta(t_{ij}) \mathbf{X}_{ij} \}, \quad t \in (0, \tau_1] \\ &= \lambda_0(t, t_{ij}) \exp \{ \beta(t_{ij}) \mathbf{X}_i(t_{ij}) \}. \end{aligned} \quad (2)$$

This is a variant of the Cox model in which both the baseline hazard function and the log hazard ratios vary with t_{ij} . We assume that $\lambda_0(t, s)$ is a smooth bivariate function defined on $t \in (0, \tau_1]$ and $s \in (0, \infty)$ (or empirically, $s \in$

$(\min\{t_{ij}\}, \max\{t_{ij}\})$, and that $\beta(s)$ is a vector of smooth functions of time. The range of t in the bivariate baseline hazard function is restricted to $(0, \tau_1]$ because all the Y_{ij} 's are no greater than τ_1 due to artificial censoring (Section 2.1). While in theory $\lambda_0(t, s)$ can be defined for $t > \tau_1$, that part of the bivariate function cannot be estimated with the landmark data set, and it is not relevant for the purpose of estimating the survival probability at τ_1 either. For the same subject i but at different landmark times t_{ij_1} and t_{ij_2} , the conditional distributions of T_{ij_1} and T_{ij_2} may be dependent.

The working assumption (2) is applied at all the measurement times $\{t_{ij}\}$. Since both $\lambda_0(t, s)$ and $\beta(s)$ are smooth functions in their arguments, we generalize (2) and assume that for any new subject whose data are identically distributed as the subjects in the landmark data set, the residual survival T at the new subject's measurement time s has the hazard function:

$$h(t \mid s, \mathbf{X}(s)) = \lambda_0(t, s) \exp\{\beta(s)\mathbf{X}(s)\}, \quad t \in (0, \tau_1]. \quad (3)$$

where $\mathbf{X}(s)$ is the vector of predictor obtained from this subject at the measurement time s , as defined in Section 2.1. Here s can be any time other than the t_{ij} 's in the training data. This generalization is not needed for model fitting but necessary to make predictions. If a new subject have survived up to time s without the event, then the predicted probability of experiencing the event before the horizon τ_1 , based on $\mathbf{X}(s)$, the data available up to prediction time s , is calculated as

$$\begin{aligned} \pi(s, \tau_1) &= P(T \in (s, s + \tau_1] \mid T > s, \mathbf{X}(s)) \\ &= 1 - \exp\left\{-\exp(\beta(s)\mathbf{X}(s)) \int_0^{\tau_1} \lambda_0(t, s) dt\right\} \end{aligned} \quad (4)$$

Note that in order to estimate the probability in (4), besides an estimator of $\beta(s)$, we only need the estimated $\lambda_0(t, s)$ function for $t \in (0, \tau_1]$.

Strictly speaking, the prediction in (4) can only be made for a subject at the measurement time, i.e., time when new data about the subject just become available. This is reasonable because when there is no new information about the subject, we stick with the old data and old prediction. If we want to predict at time s between two consecutive measurement times s_k and s_{k+1} , we propose to calculate the following predicted probability at s :

$$P(T \in (s, s_k + \tau_1] \mid T > s, s_k, \mathbf{X}(s_k)) = \frac{\pi(s_k, \tau_1) - \pi(s_k, s)}{1 - \pi(s_k, s)} \quad (5)$$

This prediction is consistent with the prediction made at s_k and properly adjusts for the fact that there is no new longitudinal information about this subject except that the event of interest did not occur between s_k and s . In doing so, we have to change the prediction horizon to $s_k + \tau_1 - s$, which is shorter than τ_1 . When $s > s_k + \tau_1$, no prediction can be made, because there are no recent data to rely on. That is an indication that the measurement frequency needs to increase. One can avoid this situation by increasing the prediction

horizon τ_1 . However, as explained in Section 2.1, increasing the prediction horizon may decrease the prediction accuracy and increase the possibility of model misspecification. Another way to avoid this situation is to choose a relatively large τ_2 and use something like “the average \mathbf{X} in the history window” as the predictor. A potential drawback of this approach is that such a predictor may not be scientifically justified to the problem under study. In chronic disease studies, the predictor variables may change slowly over time. In such situation, we can reasonably approximate $\mathbf{X}(s)$ with $\mathbf{X}(s_k)$, the most recent data in the past; in such situation, a prediction with horizon τ_1 can be made at any time without using (5).

We call model (2) a *working assumption*, and distinguish it from a *data generating assumption*. It is difficult to specify a joint distribution of the original data $\{Y_i, \delta_i, t_{ij}, \mathbf{Z}_{ij}\}$ such that the derived landmark data $\{Y_{ij}, \delta_{ij}, t_{ij}, \mathbf{X}_{ij}\}$ satisfy the Cox model exactly. Zheng and Heagerty (2005) discovered a joint distribution for the special case with a single covariate, $\tau_1 = \infty$, time-constant regression coefficients, and $t_{ij} = t'_{i'j}, \forall i \neq i'$. Simulating data under more general cases such as (3) remains a topic of further investigation. It remains a question whether such a data generating distribution does not exist in some situations. If the goal of the research is to predict the outcome, instead of exploring the causal associations among variables, it suffices to use a flexible model, such as (2), as a basis for prediction. The model must adapt well to the landmark data set, instead of the original data set. This is the rationale of the landmark approach to dynamic prediction [42, 44]. In contrast, the joint modeling approach for dynamic prediction uses data generating assumption on the original data set, but its application is more restricted for the reasons discussed in Section 1.

2.3 Model estimation

Under the working assumption (2), we estimate the unknown parameter functions $\beta(s)$ and $\lambda_0(t, s)$ by applying the kernel method to the landmark data set $\{Y_{ij}, \delta_{ij}, t_{ij}, \mathbf{X}_{ij} \mid i = 1, 2, \dots, n, j = 1, 2, \dots, n_i\}$. For estimation at time s , we assign the following kernel weight to the (i, j) -th record: $W_{ij}(s) = h^{-1}K((t_{ij} - s)/h)$, where $h > 0$ is the pre-specified bandwidth, and $K(u)$ is the kernel function. The $\beta(s)$ at time s is estimated by maximizing the following weighted partial likelihood:

$$\prod_{i=1}^n \prod_{j=1}^{n_i} \left\{ \frac{\exp[\mathbf{X}_{ij}^T \beta(s)]}{\sum_{l=1}^n \sum_{k=1}^{n_l} 1\{Y_{lk} \geq Y_{ij}\} W_{lk}(s) \exp[\mathbf{X}_{lk}^T \beta(s)]} \right\}^{\delta_{ij} W_{ij}(s)}$$

The landmark data set is a clustered survival data with possible intrasubject correlation among the n_i records from the same subject. A classical result from survival analysis is that given a marginal Cox model for multivariate survival outcomes, consistent estimators of both the log hazard ratios and the cumulative baseline hazard function can be obtained by assuming “working

independence” among data from the same cluster [17, 22, 23]. The partial likelihood above was derived from those results, with the incorporation of kernel weights. The variance of $\hat{\beta}(s)$ can be estimated either by the sandwich estimator, as described in [44], or by bootstrap, if needed.

The cumulative hazard function at prediction time s is estimated by the following Breslow estimator. Again, this is modified from the Breslow estimator for marginal Cox model with clustered data, by incorporating the kernel weights [23, 35]. Unlike the conventional case, the Breslow estimator here is a bivariate function, indexed by both t and s .

$$\hat{\Lambda}_0(t, s) = \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{1\{Y_{ij} \leq t\} \delta_{ij} W_{ij}(s)}{\sum_{l=1}^n \sum_{k=1}^{n_l} 1\{Y_{lk} \geq Y_{ij}\} W_{lk}(s) \exp[\mathbf{X}_{lk}^T \hat{\beta}(s)]}, \quad t \in (0, \tau_1]$$

We can repeat the estimation procedure above at any s , producing an estimated curve for $\beta(s)$ and an estimated bivariate surface for $\Lambda_0(t, s)$. The baseline survival function is estimated as $\hat{S}_0(t, s) = \exp\{-\hat{\Lambda}_0(t, s)\}$. For a new subject, the prediction is calculated via (4) as:

$$\hat{P}(T \in (s, s + \tau_1] | T > s, X(s)) = 1 - \exp\left\{-\exp\left(\hat{\beta}(s)\mathbf{X}(s)\right) \hat{\Lambda}_0(\tau_1, s)\right\}.$$

Given the training data, we recommend making prediction at any time $s \in [0, \max\{(Y_i - \tau_1)\delta_i\}]$, and leaving an adequate number of observed events within the prediction window $(s, s + \tau_1)$.

The construction of the landmark data set (Section 2.1) is similar to [48], except the use of a finite prediction horizon and the history window. However, the working model (2) is different. Zheng and Heagerty [48] studied a Cox model where the coefficients vary with t , the time from the landmark to the event of interest, while in this paper, the coefficients vary with s , the landmark time. Given the landmark time s and the corresponding predictors $\mathbf{X}(s)$, our Cox model essentially has time-independent covariates. Consequently, the model fitting method is also different, particularly for the baseline survival function, which is estimated nonparametrically as a bivariate function. The specification of our model (2) is similar to [42] and [44] in that the regression coefficients are smooth curves of prediction times, but the differences include that the baseline survival function in (2) is estimated as a nonparametric surface instead of using parametric functions and that the construction of the landmark data set is different (Section 2.1)

3 Double Time-dependent ROC Curve

We assess the accuracy of the predicted probability using the time-dependent Receiver Operating Characteristics (ROC) curve [13], which measures how well the predicted probabilities discriminate the subjects who experience the event before the horizon τ_1 (cases) and those who do not (controls). For a fixed prediction time, Heagerty et al [13] developed a nonparametric procedure to

estimate the bivariate distribution of the prediction probability (or biomarker) and event time, adjusting for censoring, and derive the time-dependent sensitivity and specificity from the estimated bivariate distribution. Several other methods of estimation are also available in that context [2].

In the context of dynamic prediction, the prediction time s is not fixed, and so is the evaluation of the time-dependent sensitivity and specificity. We define the dynamic time-dependent sensitivity and specificity as:

$$\begin{aligned} Se(c; s, \tau_1) &= P(Q(s) > c | s < \tilde{T}(s) \leq s + \tau_1, \tilde{Y}(s) > s, s) && \text{sensitivity} \\ Sp(c; s, \tau_1) &= P(Q(s) \leq c | \tilde{T}(s) > s + \tau_1, \tilde{Y}(s) > s, s) && \text{specificity} \end{aligned} \quad (6)$$

$\tilde{T}(s)$ is the time to event, starting from s . $Q(s)$ is the predicted probability or its logit transformation at time s . The high/low risk patient cutoff is defined at the prediction horizon τ_1 . We consider a pre-specified fixed cutoff at c for the decision rule, though c could also be specified to depend on s as $c(s)$. Let $\tilde{C}(s)$ be a random right censoring time starting from s such that $\tilde{Y}(s) = \min\{\tilde{T}(s), \tilde{C}(s), \tau_1\}$ and $\tilde{\delta}(s) = 1\{\tilde{T}(s) \leq \tilde{C}(s) \text{ and } \tilde{T}(s) \leq \tau_1\}$. Due to the ever-changing at-risk set over the prediction time s and the possible time-varying association between the prediction and the survival outcome, the time-dependent sensitivity and specificity above depend on both τ_1 and s . The existing methods for time-dependent ROC do not apply to this double time-dependence situation [2]. Therefore, we propose the following approach for estimation, which is an extension of our recent work on a nonparametric probability weighting approach to time-dependent ROC analysis [20] and R package `tdROC`. Let $\mathcal{R}(s)$ be the subjects at risk at time s , i.e., $\mathcal{R}(s) = \{i | i = 1, 2, \dots, n, \text{ and } Y_i > s\}$. For any subject $i \in \mathcal{R}(s)$, let U_i be its measurement time closest to s , i.e., $U_i = \{t_{ij} \mid |t_{ij} - s| \leq |t_{ij'} - s|, \forall j' = 1, 2, \dots, n_i\}$ (in case of ties, make a random pick). We use Q_i to denote the predicted probability or its logit transformation made at U_i , and $(\tilde{T}_i, \tilde{C}_i)$ the corresponding residual time to event and time to censoring. $\tilde{Y}_i = \min\{\tilde{T}_i, \tilde{C}_i, \tau_1\}$ and $\tilde{\delta}_i = 1\{\tilde{T}_i \leq \tilde{C}_i \text{ and } \tilde{T}_i \leq \tau_1\}$ are the observed derived time and censoring indicator.

Let $S_{\tilde{T}}(t|U_i, Q_i) = P(\tilde{T}_i \geq t|U_i, Q_i)$ be the conditional survival function of \tilde{T}_i given U_i and Q_i , evaluated at $t \in [0, \tau_1]$. We define the weight function

$$W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) = P(\tilde{T}_i \leq \tau_1 | U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) = 1 - (1 - \tilde{\delta}_i) \frac{S_{\tilde{T}}(\tau_1 | U_i, Q_i)}{S_{\tilde{T}}(\tilde{Y}_i | U_i, Q_i)} \quad (7)$$

This weight function returns the probability that subject i is a case subject, i.e., $\tilde{T}_i \leq \tau_1$, given the observed data $\tilde{Y}_i, \tilde{\delta}_i, U_i$, and Q_i . When $\tilde{Y}_i \leq \tau_1$ and $\tilde{\delta}_i = 1$, this subject is observed to be a case subject and the weight equals 1. When $\tilde{Y}_i = \tau_1$ and $\tilde{\delta}_i = 0$, this subject is observed to be a control subject and the weight equals 0. When $\tilde{Y}_i < \tau_1$ and $\tilde{\delta}_i = 0$, the status of the subject is uncertain and the weight function returns the probability of being a case subject given the observed data. That probability is calculated from the conditional survival function $S_{\tilde{T}}(\cdot)$, which can be estimated from the kernel weighted Kaplan-Meier

estimator:

$$\hat{S}_{\tilde{T}}(t|U_i, Q_i) = \prod_{\zeta \in \Omega, \zeta \leq t} \left\{ 1 - \frac{\sum_{j \in \mathcal{R}(s)} K_{h_1}(U_j, U_i) K_{h_2}(Q_j, Q_i) 1(\tilde{Y}_j = \zeta) \tilde{\delta}_j}{\sum_{j \in \mathcal{R}(s)} K_{h_1}(U_j, U_i) K_{h_2}(Q_j, Q_i) 1(\tilde{Y}_j \geq \zeta)} \right\}$$

where Ω is the set of distinct \tilde{Y}_i 's in $\mathcal{R}(s)$ with $\tilde{\delta}_i = 1$. $K_{h_1}(x, x_0) = \frac{1}{h_1} K\left(\frac{x - x_0}{h_1}\right)$

with $K(\cdot)$ being the kernel function and $h_1 > 0$ being the bandwidth. The notation is similar for $K_{h_2}(\cdot)$. The estimated weight for subject i is denoted by $\hat{W}(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i)$, with $S_{\tilde{T}}(\cdot)$ replaced by its estimator in (7).

We propose to estimate the time-dependent sensitivity at s as

$$\hat{S}e(c; s, \tau_1) = \frac{\sum_{i \in \mathcal{R}(s)} K_{h_1}(U_i, s) \hat{W}(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) 1(Q_i > c)}{\sum_{i \in \mathcal{R}(s)} K_{h_1}(U_i, s) \hat{W}(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i)} \quad (8)$$

and estimate the time-dependent specificity at s as

$$\hat{S}p(c; s, \tau_1) = \frac{\sum_{i \in \mathcal{R}(s)} K_{h_1}(U_i, s) [1 - \hat{W}(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i)] 1(Q_i \leq c)}{\sum_{i \in \mathcal{R}(s)} K_{h_1}(U_i, s) [1 - \hat{W}(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i)]} \quad (9)$$

It is important to note that no data generating or working assumptions are made to the data $\{Q_i, U_i, \tilde{T}_i, \tilde{C}_i, \tilde{Y}_i, \tilde{\delta}_i | i \in \mathcal{R}(s)\}$. In Appendix A, we prove that as $n \rightarrow \infty$, (8) and (9) converge to (6), which has the interpretation by being the time-dependent sensitivity and specificity (up to τ_1) at landmark time s . The estimation above can be applied to the training data set as well as an independent testing data set.

The time-dependent ROC curve is a plot of sensitivity (on vertical axis) vs. $1 - \text{specificity}$ (on horizontal axis). In other words, the plot consists points (x, y) , $x \in [0, 1]$ and $y \in [0, 1]$ such that $y = \widehat{ROC}(x; s, \tau_1) = \hat{S}e\left(\hat{S}p^{-1}(1 - x; s, \tau_1); s, \tau_1\right)$.

The area under the ROC curve is estimated as $\widehat{AUC}(s, \tau_1) = \int_0^1 \widehat{ROC}(x; s, \tau_1) dx$.

Similar to [29], we focus on using time-dependent ROC as a metric for prediction accuracy. Additional prospective accuracy metrics, such as the incident/static and incident/dynamic sensitivity and specificity [14], the Brier score [11], will be studied in further research. Blanche et al [3] studied a joint model of longitudinal and survival data and estimated time-dependent ROC using inverse probability censoring weighting (IPCW). Zheng and Heagerty [49] studied similar prospective accuracy metric and used semiparametric regression to characterize the bivariate distribution of the event time and biomarker values at prediction time s , possibly conditional on covariates. These researches differ from ours in both the estimation method and the context. We estimated the time-dependent ROC completely nonparametrically with a probability weight between 0 and 1, instead of the IPCW weight which is always larger than 1; the censoring distribution does not need to be estimated in our approach. The proposed time-dependent sensitivity and specificity can be calculated without estimating the bivariate distribution of the event time and biomarker (or prediction probability). We also incorporate kernel adjustment to explicitly handle irregularly spaced measurement times.

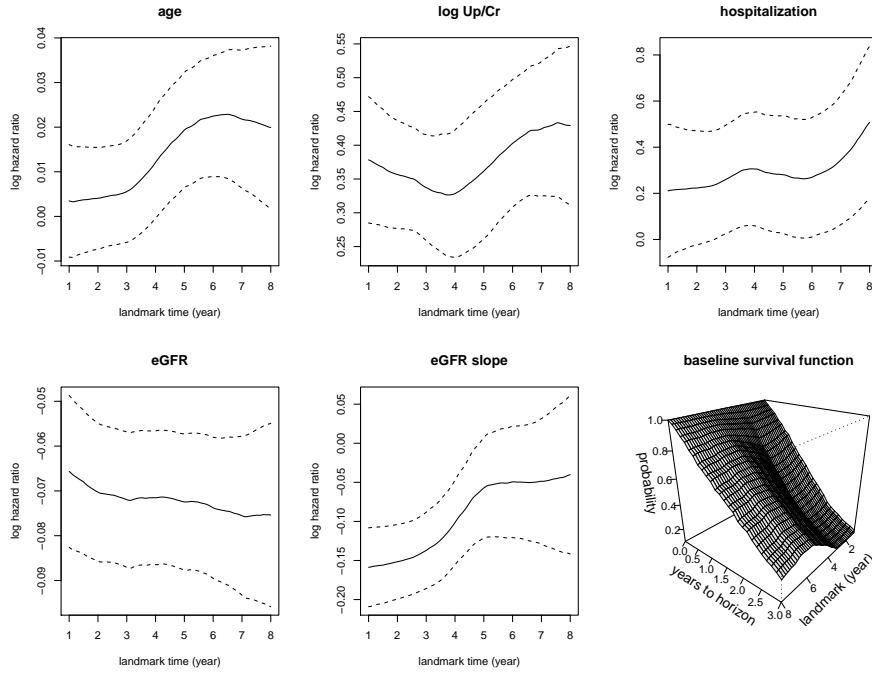


Fig. 1 The estimated log hazard ratio of each of the five predictors in a landmark model varies with the landmark time; the estimated baseline survival function also varies with the landmark time, forming a surface. The estimated log hazard ratio over landmark time was drawn with solid line and its 95% pointwise confidence interval was drawn with dashed lines. Since we predict up to a horizon of 3 years, the baseline survival function is estimated up to year 3.

4 Application

We illustrate our methodology through a data set from the African American Study of Kidney Disease and Hypertension (AASK) [10]. AASK was a multicenter randomized clinical trial of 1,094 African Americans with baseline eGFR (estimated GFR from creatinine) between 20-65 mL/min/1.73m². The study includes a trial phase and a cohort phase, with up to 12 years follow-up (median, 9). Lab and clinical data were scheduled to be collected every 6 months, but the actual times of measurements varied. ESRD, death, acute kidney injury, hospitalization, cardiovascular events, and medication history were also recorded for every patient. We illustrate with a prediction horizon of $\tau_1 = 3$ years and a history window of $\tau_2 = 3$ years. The model includes five predictors: age at time of prediction, any hospitalization in the history window, the most recent log urine protein to creatinine ratio (Up/Cr) in the history window, the eGFR at time of prediction, and the eGFR slope in the history window. The estimation of eGFR slope is described at the end of this section.

Figure 1 shows the estimated log hazard ratios of the five predictors. They vary over the landmark time in response to the ever-changing at-risk population and the possible time-varying association between the predictor and survival outcomes. The pointwise confidence intervals of Up/Cr, hospitalization, eGFR, and eGFR slope show notable deviation from 0 at most of the landmark times, suggesting their strong association with the outcome. Higher Up/Cr, presence of hospitalization, lower eGFR, and smaller eGFR slope (i.e., more steep decline) are associated with higher hazard of ESRD or death. These findings are consistent with the current clinical knowledge about the prognostic effects of these risk factors. Therefore, the landmark modeling can be used as a tool not only for generating predicted probabilities, but also for studying the association between longitudinal risk factors and clinical events. Figure 1 also shows the estimated baseline survival probability as a function of both the landmark time s and the time to event t . In the notation of equation (3), this is $\exp\left(-\int_0^t \lambda_0(u, s) du\right)$, $t \in (0, \tau_1]$. Unlike the conventional Cox model, our baseline hazard function is a bivariate function of both s and t .

Figure 2 illustrates the predicted probabilities at various clinical visits (measurement times) for six AASK subjects. The typical progression of CKD is that the eGFR declines over the course of many years with or without notable exacerbation in proteinuria, quantified by increased Up/Cr. However, the decline in eGFR or the increase in proteinuria may not be linear, as shown in this figure as well as in our previous investigation [18]. Both subject 1 and subject 2 are examples of such typical pattern. As expected, the predicted probability of ESRD or death within the horizon generally increases in response to the overall trend in the biomarkers. For subject 2, the predicted probability has a spike at a clinical visit close to Year 5. This is a response to the sharply lower eGFR measured at that visit. Subject 3's eGFR did not change much after Year 5, but the proteinuria worsened significantly after Year 7, the model responded to this change with a sharp increase in the predicted probability. Subject 4's plot illustrates that the proteinuria did not change much between Year 4 and 6, but there is a sudden drop in eGFR during that time. The model picked up this change and generated a spike in the predicted probability. We speculate that this sudden drop in eGFR may have been an Acute Kidney Injury (AKI). Note that there is a long hospitalization episode after this event. Subject 5 has a sharp drop in the eGFR at Year 2. Unlike Subject 4, the drop in eGFR is accompanied by increase in proteinuria. Both these two trends boosted the predicted probability. After the sharp decline, the eGFR stayed at approximately constant level for many years, but the Up/Cr decreased, indicating alleviation of proteinuria. The model produced decreasing predicted probabilities on this patient until the end of the follow-up, and the subject did not reach ESRD or death. For Subject 6 both biomarkers as well as the predicted probabilities were stable over the follow-up period.

We did not estimate the model parameters or calculate the predicted probabilities after Year 8 because the number of non-censored clinical events are relatively small near the end of the follow-up period. While we pre-specified

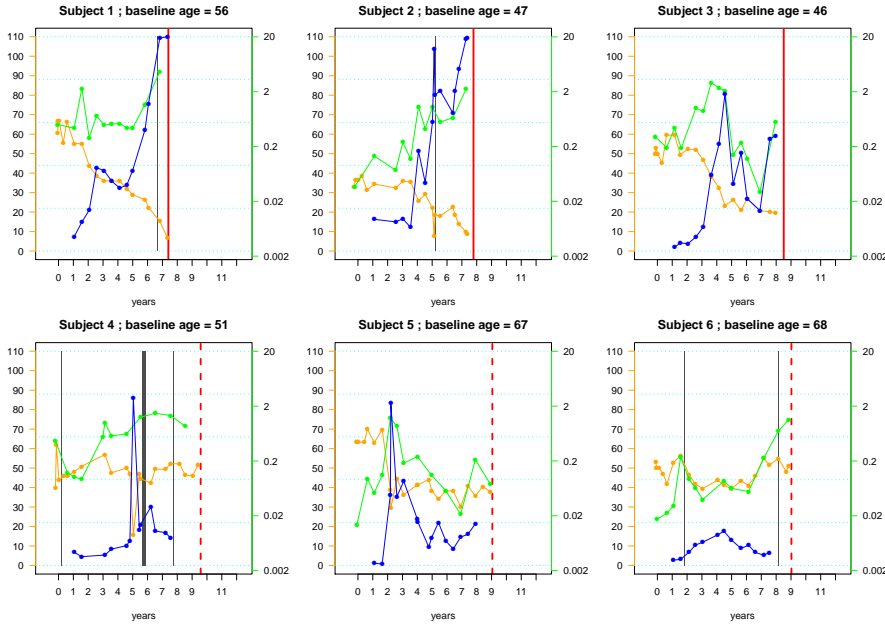


Fig. 2 An illustration of the dynamic prediction on six AASK subjects. The horizontal axis is years since randomization. The red vertical line is the time of ESRD or death (solid line) or censoring (dashed line). The orange points are eGFR ($\text{mL}/\text{min}/1.73\text{m}^2$), drawn to the axis on the left. The green points are urine protein to creatinine ratio (g/g), drawn on the log scale to the axis on the right. The blue dots are predicted probabilities of ESRD or death within a horizon of $\tau_1 = 3$ years. These predicted probabilities are calculated at each eGFR measurement time, plotted against the backdrop of six dotted light blue horizontal lines corresponding to probabilities 0, 0.2, 0.4, 0.6, 0.8 and 1 from bottom to top. The gray vertical lines represent hospitalization episodes, with the thickness of these lines proportional to the number of days in the hospital, *i.e.*, the left and right edges of the line correspond to the admission and discharge dates. The dynamic prediction probabilities were calculated using a landmark model with five predictors: age at time of prediction, any hospitalization in the past three years, the most recent log urine protein to creatinine ratio, the eGFR at time of prediction, and the eGFR slope in the past three years.

a history window size of $\tau_2 = 3$ years, that window needs to shrink at the beginning of the follow-up, when the time of prediction s is within 3 years after the baseline. This issue is discussed in Section 2.1. We did not estimate the model parameters or make prediction when $s < 1$ year in Figures 1 and 2 because there are not enough data to calculate eGFR slope as a predictor when the follow-up is less than 1 year. However, if we use only the other four predictors, the landmark model can be estimated from baseline.

The GFR is the most important measure of renal function. Hence, the typical prediction model for renal events includes eGFR as a predictor variable by default. Figure 3 shows the added value of incorporating proteinuria in the prediction model, by comparing the prediction with and without Up/Cr as a predictor. All three subjects had some spikes in Up/Cr: subject 1 at Year 2

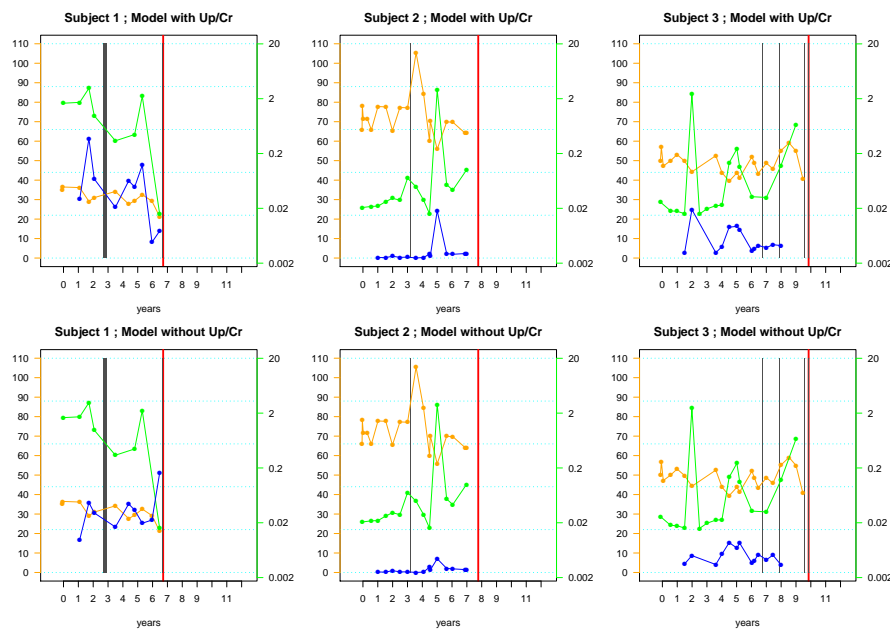


Fig. 3 An illustration of the dynamic prediction on three AASK subjects based on landmark models with or without log urine protein to creatinine ratio (Up/Cr) as a predictor. The layout and symbols of the plots are the same as those in Figure 2. The two models applied to the same subject demonstrate the difference that adding Up/Cr can make to the predicted probabilities.

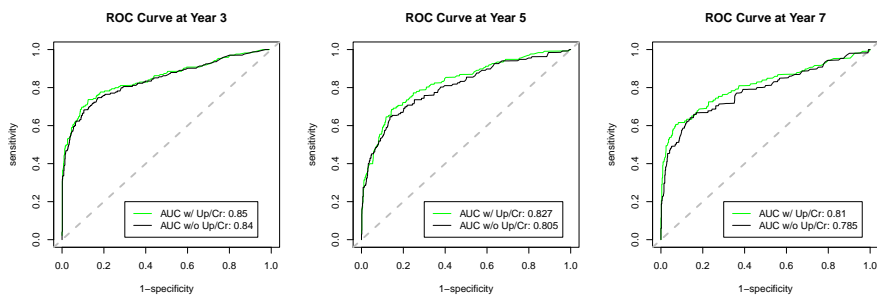


Fig. 4 The time-dependent ROC curves for prediction made at Year 3, 5, and 7. The green and black curves are ROC curves calculated from landmark models with or without using log urine protein to creatinine ratio (Up/Cr) as a predictor, respectively. The areas under the ROC curve (AUCs) are also annotated on the plots. Adding Up/Cr increases the area under the curve.

and 5, subject 2 at Year 5, and subject 3 at Year 2. The predicted probabilities from the model that incorporates Up/Cr are more responsive to the spikes.

Figures 1 and 3 show the “face validity” of the proposed dynamic prediction method. Figure 4 further demonstrates the prediction accuracy through the proposed double time-dependent ROC curves at three landmark times.

The areas under the curve (AUCs) are generally above 0.8, suggesting good discriminant power, particularly at earlier landmark time. Adding proteinuria increased the AUC, consistent with the patterns in Figures 2 and 3. We used $h_1 = 2$ years and $h_2 = 0.1$ as the bandwidth in the double time-dependent ROC calculation. The result does not appear sensitive to the bandwidths. For example, when h_1 varies between 1.5 to 3 and when h_2 alternatives between 0.1 and 0.2, the AUC at year 3 varies in a narrow range between 0.850 and 0.855 and the AUC at year 7 varies between 0.808 and 0.812. The conventional time-dependent ROC analysis [13] is a special case of the double time-dependent ROC analysis proposed in Section 3 with all the subjects having the same baseline time (i.e., $U_i \equiv U_1$). In a separate manuscript [20], we studied the performance of the proposed approach under the conventional case and demonstrate through theoretical arguments and extensive simulations that the result of the proposed weighting approach is not sensitive to the bandwidth h_2 (h_1 does not apply to the conventional case). The coefficient curves in Figure 1 were obtained using the same bandwidth h_1 as the ROC curves in Figure 4. As h_1 varies, the estimated curves in Figure 1 have different smoothness, but the overall trend does not change. The purpose of the analysis in this paper is prediction instead of estimation. Hence, one can view h_1 as a tuning parameter of the prediction algorithm and adjust it to generate the best metric of prediction accuracy in the validation data. In Figure 4 we calculated the ROC curves from the same data set that were used to develop the dynamic prediction model, which may cause the AUC to be biased higher due to the possibility of overfitting [5]. Figure 4 is used only for illustrating the proposed statistical methodology. In practice, developers of dynamic prediction models need to test the performance of the prediction in an independent data set [3] or using cross-validation.

The eGFR slope may be calculated for each subject, at each measurement time, by fitting a least squares line to the longitudinal eGFR measurements within the corresponding history window. However, when the number of eGFR measurements is small, such an approach may produce slope estimates with large variance. In the AASK analysis, we used the following procedure to calculate the eGFR slope. First, for every at-risk subject, select one history window closest to the landmark time under consideration; Second, fit a random intercept-slope model to the selected eGFR data, with proper kernel weights to adjust for the distance between the landmark time and the history window; Third, calculate the eGFR slope as the Best Linear Unbiased Prediction (BLUP). Details of this procedure and justification are described in Appendix B. This model can be fit at any landmark time, producing a “landmark” linear mixed model with all the parameters varying with the landmark time. It characterizes how the distribution of the eGFR slope vary over time with the at-risk population. The BLUP is a compromise between the individual least squares slope and the average slope of the at-risk population. When there are not enough data from the patient, the BLUP borrows information from the at risk population average. Therefore, the BLUP is numerically more stable than the least squares estimate of the individual slope.

5 Discussion

This paper proposes a general analytical framework for dynamic prediction through the landmark models. We argue that the landmark models are particularly useful in complicated problems where the other dynamic prediction approach, joint modeling, has difficulty. Our proposal uses the method of Zheng and Heagerty (2005, [48]) to construct the landmark data set, but differs from that paper in model specification and estimation, predicted probability definition and calculation, and estimating the double time-dependent ROC curves. These tasks are all unified within the proposed modeling framework by the use of kernel weighting on the scale of the landmark time. One significant advantage of the proposed methodology over the joint modeling approaches is that the computation can be easily implemented with little programming effort beyond standard statistical software. For example, the result in Figure 1 was obtained by using the `coxph()` function in the `survival` package of R with the `weight` argument. While there are only five predictors in our illustration example, the methodology can tackle large problems with many time-varying longitudinal predictors with tractable computation. Partly motivated by the AASK study application, we focus on the situation where the measurement times of longitudinal data are independent of the rest of the data and may be irregularly spaced. The methodology also applies to the special case when the measurement times are fixed at regular intervals for all subjects. The proposed framework can be further extended to replace the kernel (local constant) approach with local polynomials, though the notation will be more complicated.

The dynamic prediction and the landmark modeling are special compared to typical statistical modeling and estimation problems. Usually, statisticians hypothesize that a data generating model gives rise to the observed data, and develop methods to estimate the model parameters correctly. In this paper as well as other landmark models for dynamic prediction, there is no guarantee that the working model holds *exactly* at all the landmark times. In other words, the working model is not a data generating model, and concepts such as consistent estimation may not apply. However, from a practical perspective, the working model is useful because it can be specified flexible enough to approximate the landmark data set well, and this forms a basis for generating dynamic predictions in complicated problems where the joint modeling approach, which uses a data generating model, is difficult to implement. The lack of a data generating model makes it difficult to use simulations to study the usual properties of estimators, such as consistency and efficiency.

This paper has several limitations. First, we have assumed that the measurement times are independent of the patient's condition. This is a simplified situation. In many observational studies, it is quite possible that patients have more clinic visits when their conditions worsen, resulting in informative observational times [28, 36]. To our knowledge, no published dynamic prediction methods can deal with this challenge. This remains a topic of further research. Second, the prognostic risk factors for ESRD and death may be different. We predict their composite endpoint in this paper as an illustration of the pro-

posed modeling framework for right censored time to event outcomes. Further extension that incorporates competing risks is under investigation.

References

1. Anderson JR, Cain KC, Gelber RD. (1983) Analysis of survival by tumor response. *J Clin Oncol.*, 1(11), 710-719.
2. Blanche P, Dartigues JF, Jacqmin-Gadda H (2013). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55, 687-704.
3. Blanche P, Proust-Lima C, Loubere L, Berr C, Dartigues JF, Jacqmin-Gadda H (2015). Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1):102-113.
4. Buonaccorsi JP (1995). Prediction in the presence of measurement error: general discussion and an example predicting defoliation. *Biometrics*, 51(4), 1562-9.
5. Copas JB, Corbett P (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika* 89(2), 315-331.
6. Dabrowska DM (1987). Non-parametric regression with censored survival time data. *Scand J Statist* 14, 181-197.
7. Dafni U. (2011) Landmark analysis at the 25-year landmark point. *Circ Cardiovasc Qual Outcomes*, 4(3), 363-371.
8. Echouffo-Tcheugui JB, Kengne AP (2012). Risk models to predict chronic kidney disease and its progression: a systematic review. *PLoS Med.*, 9(11), e1001344.
9. Feuer EJ, Hankey BF, Gaynor JJ, Wesley MN, Baker SG, Meyer JS (1992). Graphical representation of survival curves associated with a binary non-reversible time dependent covariate. *Statistics in Medicine*, 11, 4554-74.
10. Gassman JJ, Greene T, Wright JT Jr, et al. Design and statistical aspects of the African American Study of Kidney Disease and Hypertension (AASK) (2003). *J Am Soc Nephrol.*, 14(7)(suppl 2), S154-S165.
11. Gerds TA, Schumacher M (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6), 1029-1040.
12. Gong Q, Schaubel DE (2013). Partly conditional estimation of the effect of a time-dependent factor in the presence of dependent censoring. *Biometrics*, 69(2), 338-347.
13. Heagerty PJ, Lumley T, Pepe MS (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56, 337-344.
14. Heagerty PJ, Zheng Y (2005) Survival model predictive accuracy and ROC curves. *Biometrics*, 61, 92-105.
15. Hsieh F, Tseng YK, Wang JL (2006). Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62(4), 1037-1043.
16. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int Suppl.* 2013, 3, 1-150.
17. Lee EW, Wei LJ, Amato DA (1992) Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: Klein JP, Goel PK (eds) *In survival analysis: state of the art.* Kluwer Academic, Norwell, Mass., pp 237-247.
18. Li L, Astor BC, Lewis J, Hu B, Appel LJ, Lipkowitz MS, Toto RD, Wang X, Wright JT Jr, Greene TH (2012) Longitudinal progression trajectory of GFR among patients with CKD. *Am J Kidney Dis.*, 59(4):504-12.
19. Li L, Greene T (2008). Varying coefficients model with measurement error. *Biometrics* 64(2):519-526.
20. Li L, Hu B, Greene T (2015) A simple method to estimate the time-dependent ROC curve under right censoring. *COBRA Preprint Series # 1167.*
21. Li L, Hu B, Greene T (2009). A semiparametric joint model for longitudinal and survival data with application to hemodialysis study. *Biometrics* 65(3):737-745.

22. Lin DY (1994) Cox regression analysis of multivariate failure time data: the marginal approach. *Stat Med.*, 13(21), 2233-2247
23. Lin DY (2007) On the Breslow estimator. *Lifetime Data Analysis*, 13, 471-480
24. Marks A, Fluck N, Black C (2014). Chronic kidney disease where next? predicting outcomes and planning care pathways. *European Medical Journal: Nephrology*, 1, 67-75.
25. National Kidney Foundation (2002). K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Am J Kidney Dis*, 39(2 Suppl 1):S1-266.
26. Parast L, Cheng SC, Cai T. (2012) Landmark prediction of long term survival incorporating short term event time information. *J Am Stat Assoc*, 107(500), 1492-1501.
27. Pepe MS (2003). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Oxford.
28. Pullenayegum EM, Lim LS (2014). Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Stat Methods Med Res*. Epub ahead of print.
29. Rizopoulos D (2011) Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67, 819-829.
30. Rizopoulos D (2012) Joint models for longitudinal and time-to-event data: with applications in R. Chapman & Hall/CRC, Boca Raton, FL, USA.
31. Rizopoulos D, Hatfield LA, Carlin BP, Takkenberg JJM (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association*, 109 (508), 1385-1397.
32. Rizopoulos D, Verbeke G, Lesaffre E (2009) Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society Series B*, 71(3), 637-654.
33. Rosansky SJ, Glasscock RJ (2014). Is a decline in estimated GFR an appropriate surrogate end point for renoprotection drug trials? *Kidney Int.*, 85(4):723-7.
34. Shiffman S. (2014) Conceptualizing analyses of ecological momentary assessment data. *Nicotine Tob Res.* 16 Suppl 2: S76-87.
35. Spiekerman CF, Lin DY (1998) Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association*, 93(443), 1164-1175
36. Sun J, Sun L, Liu D (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times. *Journal of the American Statistical Association* 102(480): 1397-1406.
37. Tangri N, Kitsios GD, Inker LA, Griffith J, Naimark DM, Walker S, Rigatto C, Uhlig K, Kent DM, Levey AS (2013). Risk prediction models for patients with chronic kidney disease: a systematic review. *Ann Intern Med*, 158(8), 596-603.
38. Taylor JM, Park Y, Ankerst DP, Proust-Lima C, Williams S, Kestin L, Bae K, Pickles T, Sandler H. (2013) Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1), 206-13.
39. Tseng YK, Hsieh F, Wang JL (2005) Joint modelling of accelerated failure time and longitudinal data. *Biometrika*, 92(3), 587-603.
40. Tsiatis AA and Davidian M (2004). An overview of joint modeling of longitudinal and time-to-event data. *Statistica Sinica*, 14, 793-818.
41. United States Renal Data System (2014) *USRDS 2014 annual data report: atlas of chronic kidney disease and end-stage renal disease in the United States*. National Institute of Diabetes and Digestive and Kidney Diseases. Bethesda, MD, USA.
42. Van Houwelingen HC (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1), 70-85.
43. Van Houwelingen HC, Putter H (2008). Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Anal.*, 14, 447-463.
44. Van Houwelingen HC, Putter H (2012) *Dynamic Prediction in Clinical Survival Analysis*. Chapman & Hall/CRC, Boca Raton, FL, USA.
45. Vickers AJ, Till C, Tangen CM, Lilja H, Thompson IM. (2011) An empirical evaluation of guidelines on prostate-specific antigen velocity in prostate cancer detection. *J Natl Cancer Inst*, 103(6):462-469.

46. Wu L, Liu W, Yi GY, Huang Y (2012) Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, Article ID 640153.
47. Zhang W, Lee S-Y, Song X (2002). Local polynomial fitting in semi-varying coefficient model. *Journal of Multivariate Analysis*, 82, 166188.
48. Zheng YY, Heagerty PJ (2005). Partly conditional survival models for longitudinal data. *Biometrics*, 61(2), 379-391.
49. Zheng Y, Heagerty PJ (2007). Prospective accuracy for longitudinal markers. *Biometrics*, 63(2), 332-41.
50. Zhou XH, Obuchowski NA, McClish DK (2011). *Statistical models in diagnostic medicine* (2nd ed). Wiley, New York.

Appendix A

We prove the consistency of the sensitivity estimator (8). The proof for the specificity estimator (9) is similar and omitted. The proof makes use of the following regularity conditions:

- (A1) The kernel function $K(\cdot)$ is a symmetric density function centered around zero with compact support.
- (A2) The variable U has bounded support; its density function $f_U(\cdot)$ is Lipschitz continuous and $f_U(s) > 0$, *a.s.*
- (A3) $h_1 \rightarrow 0$, $nh_1^2 \rightarrow \infty$, and h_1 satisfies the conditions in Lemma 1 below.

According to [6], the kernel weighted Kaplan-Meier estimator is consistent under standard regularity conditions: $\hat{S}_{\tilde{T}}(t|U_i, Q_i) = S_{\tilde{T}}(t|U_i, Q_i) + o_p(1)$. Therefore,

$$\begin{aligned} \hat{W}(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) &= 1 - (1 - \tilde{\delta}_i) \frac{S_{\tilde{T}}(\tau_1|U_i, Q_i) + o_p(1)}{S_{\tilde{T}}(\tilde{Y}_i|U_i, Q_i) + o_p(1)} \\ &= W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) + o_p(1) \end{aligned}$$

$$\begin{aligned} \hat{S}e(s, c) &= \frac{\sum_{i \in \mathcal{R}(s)} K_{h_1}(U_i, s) W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) 1(Q_i > c) + o_p(1)}{\sum_{i \in \mathcal{R}(s)} K_{h_1}(U_i, s) W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) + o_p(1)} \\ &= \frac{\sum_{i \in \mathcal{R}(s)} K_{h_1}(U_i, s) W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) 1(Q_i > c)}{\sum_{i \in \mathcal{R}(s)} K_{h_1}(U_i, s) W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i)} + o_p(1) \\ &= \frac{\sum_{i=1}^n K_{h_1}(U_i, s) W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) 1(Q_i > c) 1(\tilde{Y}_i > s)}{\sum_{i=1}^n K_{h_1}(U_i, s) W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) 1(\tilde{Y}_i > s)} + o_p(1) \end{aligned}$$

By using Lemma 2 below, the equation above equals:

$$\hat{S}e(s, c) = \frac{E \left\{ W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) 1(Q_i > c) 1(\tilde{Y}_i > s) \mid s \right\} f_U(s)}{E \left\{ W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i) 1(\tilde{Y}_i > s) \mid s \right\} f_U(s)} + o_p(1)$$

where $f_U(\cdot)$ is the density function of U_i . Replacing $W(U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i)$ by $E\left(1(\tilde{T}_i \leq \tau_1) \mid U_i, Q_i, \tilde{Y}_i, \tilde{\delta}_i\right)$ in the equation above, we have

$$\begin{aligned}\hat{S}e(s, c) &= \frac{E\left\{1(\tilde{T}_i \leq \tau_1)1(Q_i > c)1(\tilde{Y}_i > s) \mid s\right\}}{E\left\{1(\tilde{T}_i \leq \tau_1)1(\tilde{Y}_i > s) \mid s\right\}} + o_p(1) \\ &= P(Q_i > c \mid \tilde{T}_i \in (s, s + \tau_1], \tilde{Y}_i > s, s) + o_p(1)\end{aligned}$$

This completes the proof of consistency.

Lemma 1. Let $\{(X_i, Y_i), i = 1, \dots, n\}$ be independent and identically distributed random vectors, where the Y_i 's are scalar random variables. Assume that $E|Y|^s < \infty$ and $\sup_x \int |Y|^s f(X, Y) dY < \infty$, where f denotes the joint density of (X, Y) . K is a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then

$$\sup_x \left| \frac{1}{n} \sum_{i=1}^n [K_h(X_i - x)Y_i - E\{K_h(X_i - x)Y_i\}] \right| = O_p\left(\left\{\frac{\log(1/h)}{nh}\right\}^{1/2}\right)$$

provided that $n^{2\epsilon-1}h \rightarrow \infty$ for some $\epsilon < 1-s^{-1}$. (This lemma was used in [47]).

Lemma 2. Suppose $\{(Y_i, T_i), i = 1, \dots, n\}$ satisfy the conditions in Lemma 1; T_i 's satisfy the regularity conditions (A1)-(A3); $E(Y|T)$ is Lipschitz continuous. Then as $n \rightarrow \infty$:

$$n^{-1} \sum_{i=1}^n K_h(T_i - t_0)Y_i = E(Y|t_0)f(t_0) + O_p(h^2) + O_p\left(\left\{\frac{\log(1/h)}{nh}\right\}^{1/2}\right).$$

The proof of this result uses Lemma 1 and a Taylor series expansion of $E(Y|T)f(T)$ around t_0 . (This lemma was used in [19]).

Appendix B

With irregularly spaced measurement times, some subjects may not have adequate number of measurements within the history window $(s - \tau_2, s)$ for calculating the slope, or rate of change, of a biomarker. We outline a solution to this problem. Suppose we want to make a prediction on a subject at time s . The repeated biomarker measurements and the corresponding measurement times within the history window are denoted by vectors \mathbf{Z}_0 and \mathbf{t}_0 . Here we focus on the case when s equals the last measurement time in \mathbf{t}_0 . The case when s is not a measurement time has been discussed in Section 2.2. Let $\mathcal{R}(s)$ denote the set of subjects in the training data set that are at risk at time s , i.e., $\mathcal{R}(s) = \{i \mid i = 1, 2, \dots, n; Y_i > s\}$. For the i -th subject in $\mathcal{R}(s)$, let t_{ij} be the measurement time that is the closest to s , i.e., $|t_{ij} - s| \leq |t_{ij'} - s|$ for any $j' \in \{1, 2, \dots, n_i\}$ (in case of ties, make a random choice). Let \mathbf{Z}_i and \mathbf{t}_i be the corresponding vectors of biomarker measurements and measurement times in the history window (by

definition, the last element in \mathbf{t}_i is t_{ij}). Fit a random intercept and slope model to the data $\{\mathbf{Z}_i, \mathbf{t}_i, i \in \mathcal{R}(s)\}$, with the i -th subject weighted by the kernel weight $W_i = h^{-1}K((t_{ij} - s)/h)$. This is done by maximizing the following weighted multivariate Gaussian log-likelihood with respect to $\alpha(s)$, $\Sigma(s)$, and $\sigma(s)$:

$$\begin{aligned} & \sum_{i \in \mathcal{R}(s)} \left(-\frac{W_i}{2} \right) \log |\mathbf{D}_i \Sigma(s) \mathbf{D}_i^T + \sigma(s)^2 \mathbf{I}| \\ & + \sum_{i \in \mathcal{R}(s)} \left(-\frac{W_i}{2} \right) (\mathbf{Z}_i - \mathbf{D}_i \alpha(s))^T (\mathbf{D}_i \Sigma(s) \mathbf{D}_i^T + \sigma(s)^2 \mathbf{I})^{-1} (\mathbf{Z}_i - \mathbf{D}_i \alpha(s)) \end{aligned} \quad (10)$$

\mathbf{D}_i is a two-column matrix consisting of a vector of 1's and \mathbf{t}_i . $\alpha(s)$ is 2×1 vector of the fixed effect coefficients for intercept and slope. $\Sigma(s)$ is the 2×2 variance matrix of the random effects. Given the estimated parameters, the rate of change may be defined to be the second element of the following 2×1 vector:

$$\hat{\alpha}(s) + \hat{\Sigma}(s) \hat{D}_0^T \left(\mathbf{D}_0 \hat{\Sigma}(s) \mathbf{D}_0^T + \sigma(s)^2 \mathbf{I} \right)^{-1} (\mathbf{Z}_0 - \mathbf{D}_0 \hat{\alpha}(s)) \quad (11)$$

where \mathbf{D}_0 is a two-column matrix consisting of a vector of 1's and \mathbf{t}_0 . Equation (11) is the best linear unbiased prediction of the random intercept and slope. Since $\alpha(s)$, $\Sigma(s)$, and $\sigma(s)$ can be estimated at any landmark time s , their trajectories characterize the temporal change of the at-risk population.

We emphasize that the procedure above is not intended to provide a consistent or unbiased estimate of the slope of the biomarker trajectory at time s for each subject, or the mean slope of the at-risk subjects at time s , because the trajectory may be nonlinear and τ_2 does not go to 0 as n increases. The procedure should be viewed as a principled way of defining biomarker slope other than using the least squares estimate. It causes the individual least squares slope to shrink toward the mean slope of the at-risk subjects, and thus increases the stability of individual slopes. When the individual biomarker trajectory is nonlinear and smooth, estimating the slope as a smooth time-varying function is difficult in both the landmark modeling and joint modeling frameworks, unless strong assumptions are made for the trajectory shape or large numbers of repeated measures per subject are available. In clinical practice, the concept of a time-varying first-order derivative of a nonlinear smooth "true" trajectory is rarely used in prognostic evaluation. Physicians often use the least square slope in the history window to describe, in broad strokes, the overall trend of the biomarker during that time, even when in reality most of the biomarker trajectories are nonlinear to certain extent and their first order derivatives are never constant within the history window. Examples include the prostate specific antigen (PSA) velocity [45] and GFR slope [33]. The procedure above is an algorithmic formulation of the concept of rate of change used in clinical practice. Therefore, this linear mixed model with time-varying parameters should be viewed as a working assumption.

Even when the “true” individual trajectory is indeed linear within the history window with a constant underlying “true” slope, the BLUP is not the true slope and their difference is often viewed as a manifestation of measurement error in the biomarker [21]. The magnitude of the error is of order $\sigma_i/\sqrt{m_i}$, where σ_i is the residual variance of subject i and m_i is the number of eGFR measurements within the history window. The measurement error model literature showed that covariate measurement error does not cause bias in prediction if the distribution of the observed data is the same in the training and testing data sets [4]. Therefore, despite the measurement error, we can still use the BLUP for prediction purpose. This is the justification for the proposed biomarker slope definition. However, the assumption that the training and testing data sets have the same distribution implies that the distribution of the measurement times t_{ij} must also be the same, as the measurement times are treated as random in this paper (Section 2.1). If, for example, we develop a landmark model out of a training data set where every subject follows a six month visit schedule with some random deviation, and apply it to a new data set where every patient follows a three month visit schedule, then it would not be appropriate to use the BLUP with a time-varying linear mixed model developed from the training data.