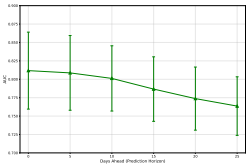
Two-Stage Modeling for Dynamic Survival Prediction from Longitudinal Data

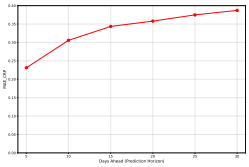
Anonymous Author(s)
Affiliation
Address
email

Abstract

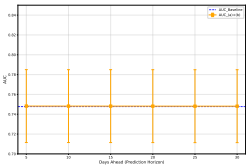
In dynamic survival prediction, landmarking predicts risk at a fixed future horizon from data observed at a single landmark time. Accuracy typically worsens as the prediction horizon increases. We propose a simple yet novel two-stage extension: first forecast near-future *laboratory measurements*, then predict the outcome over the resulting shorter window. This *naive* two-stage modeling already improves performance; an *extended* version that also passes distributional summaries from the measurement forecast (e.g., predictive mean and variance) to the outcome model yields further gains. In experiments on a hospital cohort with routine laboratory measurements and the MIMIC-IV dataset, the two-stage approach consistently outperforms one-stage landmarking across horizons, with the extended variant best overall. In aggregate, our method improves AUC by about 3 percentage points at most compared with the one-stage baseline.



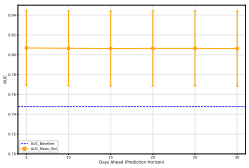
(a) Predictive accuracy of one-stage conventional landmarking as a function of the horizon d (x-axis: d in days; y-axis: dynamic prediction accuracy (AUC)).



(b) Error in forecasting the Stage 1 biomarker k days ahead in the two-stage modeling (x-axis: k in days; y-axis: Stage 1 forecasting error (MAE)).



(c) Predictive accuracy of naïve two-stage modeling as a function of k (x-axis: k in days; y-axis: dynamic prediction accuracy (AUC)).



(d) Predictive accuracy of extended two-stage modeling as a function of k (conditioning on the Stage 1 predictive mean and standard deviation): (x-axis: k in days; y-axis: dynamic prediction accuracy (AUC)).

Figure 1: Results on a hospital cohort from our institution. For clarity, all panels use only a single biomarker (C-reactive protein (CRP)). In the main experiments, we use four laboratory tests, which improves overall accuracy but attenuates the clear monotonic trends visible here.

1 Introduction

Prognostic prediction is fundamental in medicine: it supports patient decision-making, informs clinicians in choosing optimal treatments, and enables early identification of high-risk patients to catalyze new therapy development [1, 2]. In routine clinical practice, however, a patient's condition evolves over time. Conventional prognostic models typically issue a prediction from a single snapshot

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

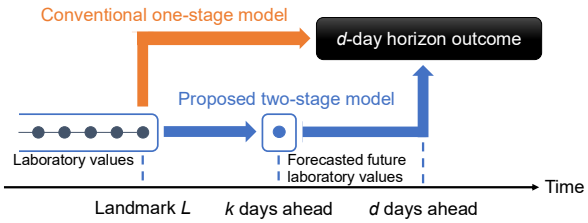


Figure 2: Overview of the baseline one-stage model and proposed two-stage model.

of patient status and cannot readily adapt to subsequent changes. This limitation is particularly consequential in intensive care and oncology, where patient trajectories are driven by rapid physiologic fluctuations and treatment effects; in such settings, predictions must be revised as new information arrives. Dynamic prediction models address this need by updating risk estimates over time using longitudinal data [3]. From a statistical perspective, landmarking [4, 5] and joint modeling of longitudinal and time-to-event data [6] are established paradigms. In parallel, the machine learning community has developed models that leverage longitudinal laboratory data for survival prediction [7, 8], and recent deep learning approaches offer flexible function classes for continuous-time dynamic prediction [9, 10]. Empirical studies demonstrate clinical utility in intensive care [11] and allogeneic hematopoietic cell transplantation (allo-HCT) [12, 13].

In this paper, we aim to improve predictive accuracy in landmarking. We first describe the conventional formulation and its drawback, and then extend it to a two-stage framework (see Fig. 2). Let L denote the landmark time at which a prediction is issued. Let T be the event time of interest. Let $R := T - L$ denote the (random) residual time-to-event measured from the landmark. Let $\{X_t\}_{t \geq 0}$ be a vector-valued longitudinal laboratory measurement process, and write the laboratory measurement history up to (and including) L as $\mathcal{X}_L := \{X_s : s \leq L\}$. Given a prediction horizon d (e.g., $d = 30$ days), conventional landmarking uses $\mathcal{X}_L$ to predict the d-day horizon outcome $Y_{L,d} = \mathbb{1}\{R \leq d\}$[1]. However, as illustrated in Fig. 1(a), predictive accuracy tends to deteriorate as the target day moves farther from the landmark.

To improve accuracy, we consider *shortening the effective prediction horizon*. The intuition is simple. If the laboratory value on the target day (i.e., $L + d$) were available, we would condition on that value and predict the same-day outcome, which would reduce error. Of course, those future laboratory measurements are unknown at the landmark. We therefore insert an intermediate step that forecasts a near-future laboratory value and then predicts the outcome over a shorter interval, which we refer to as *two-stage modeling*. Choose an offset $k \in [0, d]$ [2]. **Stage 1** forecasts the laboratory value(s) at time $L + k$ from the information available at L, producing $\widehat{X}_{L+k}$. **Stage 2** uses $\widehat{X}_{L+k}$ to predict the outcome over the shorter horizon $d - k$. As shown in Fig. 1(b), increasing k generally laboratory value forecasting, but the survival horizon $d - k$ becomes shorter, which mitigates prediction error relative to the one-stage baseline. In Fig. 1(c), the variant that uses only the Stage 1 point forecast $\widehat{X}_{L+k}$ slightly outperforms the one-stage baseline in aggregate; we refer to this as *naive two-stage modeling*. Furthermore, incorporating distributional summaries of the Stage 1 predictions (e.g., the mean and variance of the predictive distribution for X_{L+k}) into Stage 2 yields further gains in predictive accuracy (Fig. 1(d)); we refer to this as *extended two-stage modeling*.

In the remainder of the paper, we demonstrate the effectiveness of our simple yet novel method and identify avenues for further improvement.

2 Methods

Stage 1 We constructed separate predictive models for each of the four laboratory variables. In each model, the laboratory value at the landmark was used as an explanatory variable of a regression model. The distribution of the laboratory value of k days ahead was predicted using Bayesian linear regression. Since the outputs of Stage 1 were subsequently used in Stage 2, Bayesian linear regression was employed to obtain predictive distributions of future laboratory values, thereby accounting for

[1] For exposition, we ignore time-invariant baseline covariates.

[2] $k = 0$ recovers the standard one-stage landmarking formulation.

prediction uncertainty. The predictive distributions were estimated by stochastic variational inference. From these distributions, various statistics such as the mean and standard deviation can be obtained, and we experimentally evaluated which of these features are most informative for predicting survival. While the basic setting involved forecasting 30 days ahead, we also examined multiple prediction horizons of $k = 5, 10, 15, 20, 25, 30$ days in the experiments.

Stage 2 We used age at transplantation and rDRI, together with the mean and standard deviation obtained from 2,000 samples drawn from the predictive distributions of 30-day laboratory values estimated in Stage 1, as explanatory variables to predict 30-day mortality from each prediction point. As the prediction model, we employed a dynamic survival prediction framework that combines landmarking with the Cox proportional hazards model. A landmark represents the time point at which prediction is made, and multiple landmarks were set to construct regression models at each prediction time.

3 Experiments

3.1 Experimental Settings

We retrospectively analyzed patients with hematologic malignancies who underwent allogeneic hematopoietic stem cell transplantation at our institute (anonymized for double-blind review) between January 2000 and December 2020. As pre-transplant factors, we considered age at transplantation and the refined Disease Risk Index (rDRI), which stratifies disease and disease status. As post-transplant longitudinal variables, we used serum albumin, C-reactive protein (CRP), lactate dehydrogenase (LDH), and platelet count. In total, data from 519 transplant recipients were included, with a median age of 48 years (range, 17—72). The median follow-up period was 677 days (range, 27—5510), and the 1-year overall survival rate was 66.5% (95% CI, 62.6–70.7%).

Also, to further evaluate the accuracy of the proposed method, we conducted additional experiments using the MIMIC-IV v2.2 (Medical Information Mart for Intensive Care IV) database, which contains data from 299,712 intensive care unit (ICU) patients at the Beth Israel Deaconess Medical Center [14]. This dataset includes not only admission information but also vital signs and laboratory test results obtained during ICU stays, making it suitable for our task as it provides longitudinal patient data with survival labels. We excluded patients younger than 18 years and older than 90 years, as well as those who had not stayed in the ICU for at least 48 hours. After these criteria, 21,677 patients remained. As static variables, we used age and gender. For time-series variables, we used heart rate (HR), respiratory rate (RR), body temperature (°F), oxygen saturation (SaO2), and arterial blood pressure (systolic: ABPs, diastolic: ABPd). In this dataset, we defined 24-hour mortality as the prediction outcome.

Experimental Task For the our-institute dataset, we compare the predictive accuracy of the proposed method and the baseline method for 30-day mortality. In this setting, we also examine the first stage by varying the forecasting horizon of laboratory values to 5, 10, 15, 20, 25, and 30 days. For the MIMIC dataset, we focus on 24-hour mortality prediction to align the prediction horizon with the outcome definition. In addition, to investigate the effect of dataset size, we performed bootstrap sampling of patients while fixing the test set with a random seed, and examined the relationship between the number of patients and predictive accuracy.

Evaluation metrics We conducted a 10-fold cross-validation to evaluate the model's generalization performance. The data were split into 10 folds on a per-patient basis to ensure that the same patient was not included in multiple folds. In each iteration, one fold was used as the test set, and the remaining nine folds were used as the training set. A subset of the training data was further split into a validation set. The final performance was obtained by aggregating the results across the 10 test folds. For accuracy assessment, we used time-dependent ROC curves and reported the average AUC across the ten folds as the evaluation metric.

3.2 Results

Table 1 shows the AUC values when varying the prediction horizon of laboratory values in Stage 1 to 5, 10, 15, 20, 25, and 30 days. While the baseline method achieved an AUC of 0.8730, the

Table 1: Comparison between baseline one-stage model and proposed two-stage model.

Days ahead	Baseline one-stage model (AUC)	Proposed two-stage model (AUC)
5	0.8730	**0.9022**
10		**0.9025**
15		**0.9027**
20		**0.9031**
25		**0.9032**
30		**0.9035**

Table 2: Comparison of AUC between methods on the our-institute and MIMIC datasets

Method	our-institute (AUC)	MIMIC (AUC)
Baseline one-stage model	0.8730	0.6966
Proposed two-stage model	**0.9035**	**0.7105**

proposed method exhibited a slight improvement as the prediction horizon increased, reaching the highest value of 0.9035 at 30 days. These results indicate that the proposed method consistently maintains high performance across horizons, with the 30-day horizon being the most effective. Table 2 presents the comparison between the proposed method and the baseline method under the 30-day horizon condition, which yielded the best results for the proposed method. In the our-institute dataset, the proposed method achieved an AUC of 0.9035, compared to 0.8730 for the baseline method, representing an improvement of 0.0305. In the MIMIC-IV dataset, the proposed method also outperformed the baseline, with an AUC of 0.7105 versus 0.6966. These results demonstrate that the proposed method consistently achieves higher accuracy than the baseline across different datasets. Figure 4 further illustrates the results of the additional experiment investigating the relationship between dataset size and predictive accuracy. The results show that the proposed method maintains its advantage over the baseline across different patient sample sizes, although the degree of improvement varies with the number of patients.

4 Conclusion

We presented a simple extension of landmarking for dynamic survival prediction that mitigates the degradation with increasing prediction horizon. The core idea is to shorten the effective horizon by first forecasting near-future laboratory data and then predicting the outcome over the shorter window. We studied two instantiations: a *naive* variant that conditions on the Stage 1 point forecast and an *extended* variant that additionally propagates distributional summaries from Stage 1. On an institutional hospital cohort with routine laboratory data and the MIMIC-IV dataset, both variants consistently outperformed one-stage landmarking across horizons, with the extended approach performing best; overall, discrimination improved by roughly 3.1 and 1.4 percentage points in AUC relative to the one-stage baseline, respectively.

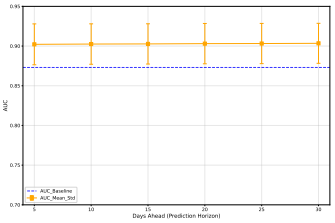
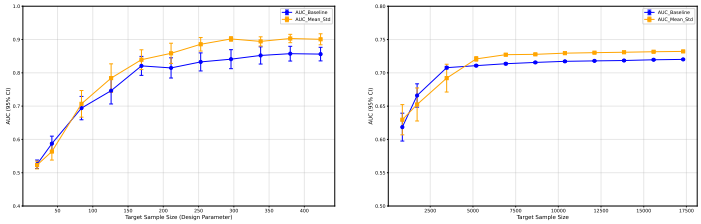


Figure 3: Comparison of AUC between the baseline one-stage model (AUC = 0.8730) and the proposed two-stage model (AUC = 0.9035).



(a) Our-institute dataset (b) MIMIC-IV dataset

Figure 4: Relationship between patient sample size and predictive accuracy for the baseline one-stage model and proposed two-stage model.

References

[1] K. G. M. Moons, D. G. Altman, J. B. Reitsma, J. P. A. Ioannidis, P. Macaskill, E. W. Steyerberg, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. *Annals of Internal Medicine*, 162(1):W1–W73, 01 2015. doi: 10.7326/M14-0698.

[2] Maarten van Smeden, Johannes B. Reitsma, Richard D. Riley, Gary S. Collins, and Karel G. M. Moons. Clinical prediction models: diagnosis versus prognosis. *Journal of Clinical Epidemiology*, 132:142–145, 04 2021. doi: 10.1016/j.jclinepi.2020.12.001.

[3] David A. Jenkins, Matthew Sperrin, Glen P. Martin, and Niels Peek. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic and Prognostic Research*, 2(1):23, 2018. doi: 10.1186/s41512-018-0033-4.

[4] Hans C. van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007. doi: 10.1111/j.1467-9469.2006.00529.x.

[5] Hans C. van Houwelingen and Hein Putter. Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Analysis*, 14(4):447–463, 2008. doi: 10.1007/s10985-008-9099-5.

[6] Dimitris Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829, 09 2011. doi: 10.1111/j.1541-0420.2010.01546.x.

[7] Kristin L. Pickett, Karthik Suresh, Kenneth R. Campbell, Scott Davis, and Elizabeth Juarez-Colunga. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC Medical Research Methodology*, 21(1):216, 2021. doi: 10.1186/s12874-021-01367-8.

[8] Jinzhu Lin, Kai Li, and Sheng Luo. Functional survival forests for multivariate longitudinal outcomes: Dynamic prediction of alzheimer's disease progression. *Statistical Methods in Medical Research*, 30(1):99–111, 2021. doi: 10.1177/0962280220909251.

[9] Mirsad Mesinovic, Peter Watkinson, and Tingting Zhu. Dysurv: dynamic deep learning model for survival analysis with conditional variational inference. *Journal of the American Medical Informatics Association*, 2024. doi: 10.1093/jamia/ocae271. Article ID: ocae271.

[10] Liang Zeng, Jing Zhang, Wei Chen, and Yiming Ding. tdcoxsnn: Time-dependent cox survival neural network for continuous-time dynamic prediction. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 74(1):187–203, 2025. doi: 10.1093/jrsssc/qkac050.

[11] Linda Lapp, Marc Roper, Kimberley Kavanagh, Matt-Mouley Bouamrane, and Stefan Schraag. Dynamic prediction of patient outcomes in the intensive care unit: A scoping review of the state-of-the-art. *Journal of Intensive Care Medicine*, 2023. doi: 10.1177/08850666231166349. URL `https://journals.sagepub.com/doi/10.1177/08850666231166349`.

[12] Soojin Kim, Brent Logan, Mary Riches, Mei-Jie Chen, and Kwang Woo Ahn. Statistical methods for time-dependent variables in hematopoietic cell transplantation studies. *Transplantation and Cellular Therapy*, 27(2):125–132, 02 2021. doi: 10.1016/j.jtct.2020.09.030.

[13] Shota Nakako, Hideaki Okamura, Iori Yokota, Yusuke Umemoto, Masaru Horiuchi, Keisuke Sakatoku, et al. Dynamic relapse prediction by peripheral blood wt1 mrna after allogeneic hematopoietic cell transplantation for myeloid neoplasms. *Transplantation and Cellular Therapy*, 30(11):1088.e1–1088.e12, 2024. doi: 10.1016/j.jtct.2024.09.004.

[14] Johnson A.E.W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x. URL `https://doi.org/10.1038/s41597-022-01899-x`.