Two-Stage Modeling for Dynamic Survival Prediction from Longitudinal Data

Tomoya Ogawa
Osaka Metropolitan University

Masakazu Iwamura
Osaka Metropolitan University

Soichiro Nakako
Osaka Metropolitan University

Hiroshi Okamura
Osaka Metropolitan University

Akinori Nishikawa
Wakayama Medical College

Koichi Kise
Osaka Metropolitan University

Abstract

Predicting a patient's future health risk is essential for medical care, especially when deciding treatments or identifying patients who may need urgent attention. However, many patients' conditions change over time, meaning that predictions made at one point may become outdated. Dynamic prediction methods address this challenge by updating risk estimates as new laboratory measurements become available. A widely used method called landmarking predicts the risk of an event (such as relapse or death) based on clinical data observed at a specific time point. However, its accuracy often becomes worse as the prediction window becomes longer—that is, the farther into the future we try to predict, the less reliable the prediction becomes. To address this limitation, we propose a simple but effective extension of landmarking called a two-stage modeling approach. Instead of predicting the outcome far into the future all at once, our method first forecasts near-future laboratory measurements and then uses those predicted values to make a shorter-range survival prediction. Even this straightforward *naive* two-stage method improves accuracy compared with standard landmarking. We further extend this approach by also using summary information about the uncertainty of the laboratory forecast (such as predicted mean and variance). This *extended* two-stage model achieves better. Across multiple clinical datasets, the proposed two-stage approach consistently improves prediction accuracy over conventional landmarking. These results show that a simple forecasting step can meaningfully improve dynamic risk prediction, offering a practical direction for clinical machine learning models that use longitudinal laboratory values.

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 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

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

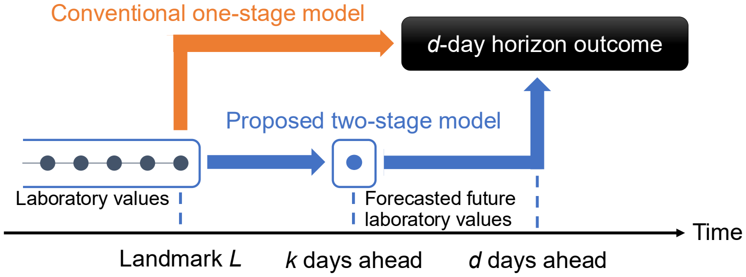
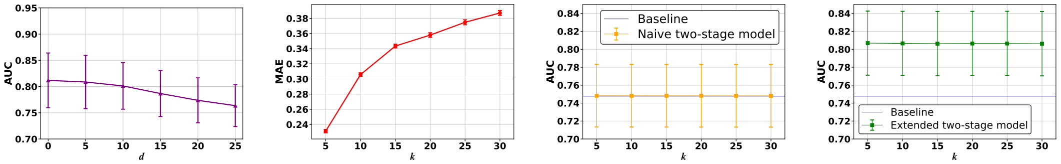


Figure 1: Overview of the baseline one-stage model and the proposed two-stage model.



(a) Predictive accuracy of conventional (baseline) one-stage method 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 laboratory value k days ahead in the two-stage modeling (x-axis: k in days; y-axis: Stage 1 forecasting error (MAE)).

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

(d) Predictive accuracy of the extended two-stage method as a function of k: (x-axis: k in days; y-axis: dynamic prediction accuracy (AUC)).

Figure 2: Results on the OMU dataset. For clarity, all panels use only a single laboratory value (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. Error bars represent 95% confidence intervals (CI). See Table 2 and Fig. 3 for the result on all laboratory variables.

longitudinal data [3]. From a statistical perspective, landmarking [4, 5] and joint modeling of longitudinal and time-to-event data [6] are established paradigms, and further extensions such as Landmarking 2.0 [7] have also been proposed. In parallel, the machine learning community has developed models that leverage longitudinal laboratory values for survival prediction [8, 9], and recent deep learning approaches offer flexible function classes for continuous-time dynamic prediction [10, 11]. Empirical studies demonstrate clinical utility in intensive care [12] and allogeneic hematopoietic cell transplantation (allo-HCT) [13, 14].

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. 1). 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 $Y_{L,d} = \mathbb{I}\{R \leq d\}$[1]. However, as illustrated in Fig. 2(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 values and then predicts the outcome over a shorter interval, which we refer to as *two-stage model*. 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. 2(b), increasing k generally laboratory

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

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

value forecasting, but the survival horizon $d - k$ becomes shorter, which mitigates prediction error relative to the one-stage baseline. In Fig. 2(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 model*. 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. 2(d)); we refer to this as *extended two-stage model*.

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

2 Methods

The proposed two-stage model predicts the outcome via the following two stages.

Stage 1 A separate estimation model is constructed for each laboratory variable. In each model, the laboratory value(s) observed at (or up to) the landmark time are used as explanatory variables, and Bayesian linear regression is applied. During training, Bayesian linear regression estimates the posterior distributions of the regression coefficients. Missing laboratory measurements are imputed using the last observation carried forward (LOCF) method [15]. During inference, a future laboratory value is estimated using the regression coefficients randomly drawn from their estimated distributions. We can use the predicted laboratory value of k days ahead in Stage 2. However, to enhance the performance, we use their distributional summary (i.e., mean and standard deviation) calculated from the estimated values over S sampling trials.

Stage 2 Stage 2 uses a Cox proportional hazards model, as in the baseline one-stage framework. In the baseline one-stage model, the laboratory value(s) at (or up to) the landmark time are used to predict d-day mortality. However, in the proposed two-stage model, the distributional summary calculated in Stage 1.

3 Experiments

3.1 Experimental Settings

Datasets We used two datasets: Osaka Metropolitan University (OMU) and MIMIC-IV datasets.

The first dataset consists of a retrospective cohort of patients with hematologic malignancies who underwent allogeneic hematopoietic stem cell transplantation at OMU Hospital 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 one-year overall survival rate was 66.5% (95% CI, 62.6–70.7%).

The second dataset was derived from the MIMIC-IV (Medical Information Mart for Intensive Care IV) v2.2 database, which contains data from 299,712 intensive care unit (ICU) patients at the Beth Israel Deaconess Medical Center [16]. 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), as these variables exhibited similar acquisition frequencies in MIMIC-IV.

Experimental Tasks For the OMU dataset, we evaluated the models based on their predictive accuracy for 30-day mortality (i.e., $d = 30$). For the MIMIC-IV dataset, the prediction task was 24-hour mortality (i.e., $d = 1$). For both datasets, the number of sampling trials was set to $S = 1$ for the naive two-stage model and $S = 2000$ for the extended version.

Table 1: Comparison of the baseline one-stage model and the proposed two-stage model on the OMU and MIMIC-IV datasets, with AUC reported as mean ± SD. Best results are shown in bold. Corresponding p-values are also provided, and results with $p < 0.05$ are marked with an asterisk to indicate statistical significance.

	OMU	MIMIC-IV
Baseline one-stage model	0.8730 ± 0.0630	0.6966 ± 0.0050
Proposed two-stage model	**0.9035** ± 0.0334	**0.7105** ± 0.0021
p-value	0.021*	0.031*

Table 2: Comparison of joint modeling [6, 17], the baseline one-stage model, and the proposed two-stage model on the OMU dataset across different values of k, with AUC reported as mean ± SD. Best results are shown in bold.

k	Joint modeling [6, 17]	Baseline one-stage model	Proposed two-stage model
5	0.8360 ± 0.0854	0.8730 ± 0.0630	**0.9022** ± 0.0341
10			**0.9025** ± 0.0338
15			**0.9027** ± 0.0336
20			**0.9031** ± 0.0337
25			**0.9032** ± 0.0336
30			**0.9035** ± 0.0334

We conducted K-fold cross-validation to evaluate the generalization performance of the models. $K = 10$ for the OMU dataset and $K = 5$ for the MIMIC-IV dataset. The data were split into K folds on a per-patient basis to ensure that the same patient was not included in multiple folds; one fold was used as the test set, and the remaining $K - 1$ folds were used as the training set. After splitting the dataset into training and test sets, 20% of the training data was further allocated as a validation set and used for early stopping and learning rate scheduling in the first-stage regression. For accuracy assessment, we used time-dependent ROC curves and reported the average AUC along with its standard deviation across the K folds as the evaluation metric.

3.2 Results

Table 1 presents the comparison between the proposed two-stage method and the baseline one-stage method. For the OMU dataset, the proposed method achieved an AUC of 0.9035, compared to 0.8730 for the baseline method, representing an improvement of 0.0305. For the MIMIC-IV dataset, the proposed method also outperformed the baseline, with an AUC of 0.7105 versus 0.6966, yielding an improvement of 0.0139. These results demonstrate that the proposed method consistently achieves higher predictive accuracy across datasets. The one-sided exact paired permutation test confirmed that these differences were statistically significant, with p-values of 0.021 for the OMU dataset and 0.031 for the MIMIC-IV dataset. With the significance level set at $p = 0.05$, both p-values fall below this criterion, indicating that the results are statistically significant.

Table 2 and Fig. 3 show the AUC values for the OMU dataset. Table 2 compares three approaches: joint modeling [6, 17], the baseline one-stage model, and the proposed two-stage model. Fig. 3 compares different three approaches: the baseline one-stage model, the naive two-stage model, and the extended two-stage model (proposed method). For the two-stage models, the prediction horizon for laboratory values in Stage 1 (i.e., k) was varied across 5, 10, 15, 20, 25, and 30 days. Across all values of k, the extended two-stage model (proposed method) consistently achieved the highest AUC, followed by the naive two-stage model, the baseline one-stage model, and joint modeling. As k increased, the predictive accuracy monotonically increased, with the extended two-stage model reaching the highest AUC of 0.9035 at a prediction horizon of 30 days.

Fig. 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.

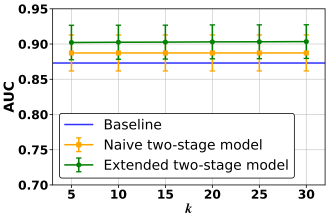
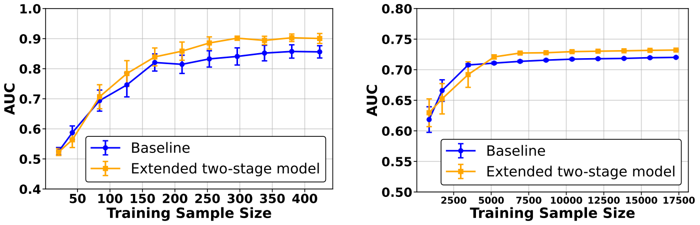


Figure 3: Comparison of the baseline one-stage model, the naive two-stage model, and the extended two-stage model (proposed), measured by AUC. Error bars correspond to the 95% CI.



(a) OMU 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. Error bars correspond to the 95% CI.

4 Conclusion

We presented a simple extension of landmarking for dynamic survival prediction that mitigates performance degradation as the prediction horizon increases. The core idea is to shorten the effective prediction window by first forecasting near-future laboratory measurements and then predicting the outcome over this shorter interval. We explored two variants of this approach: a *naive* two-stage method that conditions only on the point estimates from Stage 1, and an *extended* two-stage method that additionally propagates distributional summaries (i.e., the mean and standard deviation) from Stage 1. Across experiments, the extended two-stage method (proposed method) consistently outperformed the baseline one-stage method in terms of predictive accuracy measured by AUC, yielding improvements of 0.0305 on the OMU dataset and 0.0139 on the MIMIC-IV dataset. Future work includes extending the evaluation to comparisons with Landmarking 2.0 and replacing the Cox proportional hazards model with nonlinear predictive models.

Acknowledgments and Disclosure of Funding

This work was supported by JST ASPIRE Program JPMJAP2403.

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] Hein Putter and Hans C. van Houwelingen. Landmarking 2.0: Bridging the gap between joint models and landmarking. *Statistics in Medicine*, 41(11):1901–1917, 2022. doi: https://doi.org/10.1002/sim.9336. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9336`.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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`.

[13] 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.

[14] Soichiro Nakako, Hiroshi Okamura, Isao Yokota, Yukari Umemoto, Mirei Horiuchi, Kazuki Sakatoku, et al. Dynamic relapse prediction by peripheral blood wt1mrna 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.

[15] Dennis B. Gillings and Gary G. Koch. The application of the principle of intention-to-treat to the analysis of clinical trials. *Therapeutic Innovation & Regulatory Science*, 25:411 – 424, 1991. URL `https://api.semanticscholar.org/CorpusID:74277107`.

[16] 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`.

[17] Dimitris Rizopoulos, Pedro Miranda Afonso, and Grigorios Papageorgiou. *JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data*, 2025. URL `https://github.com/drizopoulos/jmbayes2`. R package version 0.5-97.

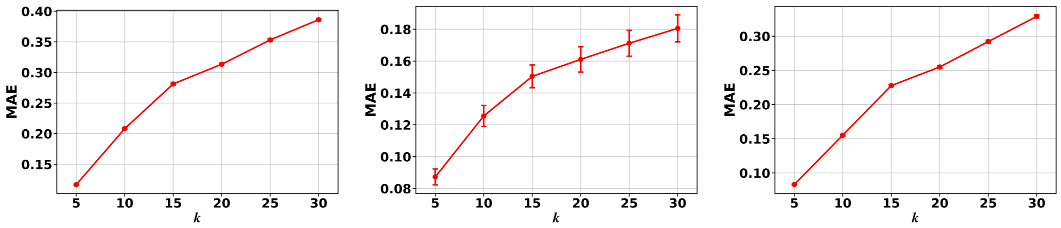
[18] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches, 2018. URL `https://arxiv.org/abs/1803.04386`.

[19] Ranganath Krishnan, Pi Esposito, and Mahesh Subedar. Bayesian-torch: Bayesian neural network layers for uncertainty estimation. `https://github.com/IntelLabs/bayesian-torch`, January 2022. URL `https://doi.org/10.5281/zenodo.5908307`.

A Implementation Details

A.1 Joint Modeling

We applied a joint modeling approach as a baseline to compare with our proposed two-stage method. It was fitted to the OMU dataset and used to dynamically predict 30-day mortality, implemented using



(a) Albumin

(b) LDH

(c) Platelet Count

Figure 5: Individual MAE curves for Albumin, LDH, and Platelet Count.

the JMbayes2 package [17]. The model consists of two linked submodels: (1) linear mixed-effects models for four longitudinal values (CRP, albumin, LDH, and platelet count) with patient-specific random intercepts and slopes; (2) a Cox proportional hazards model associated with the laboratory value trajectories through current value and slope parameterizations. Similar to the landmark approach, we employed 10-fold cross-validation at the patient level, using LOCF for missing values, and generated dynamic predictions at the same landmark time points for 30-day mortality risk. Predictive performance was evaluated using time-dependent ROC curves and reported the average AUC across the 10 folds.

A.2 Bayesian linear regression

We employed Bayesian linear regression with variational inference using the Flipout method [18], implemented with the Bayesian-Torch library [19]. The model was optimized by minimizing MSE loss combined with KL divergence, with Gaussian priors N(0, 1) for weights. Predictive uncertainty was estimated via 2,000 Monte Carlo samples.

B Additional MAE Results for Other Laboratory Values in the OMU Dataset

Fig. 2(b) presented the Stage 1 MAE curve for CRP. Fig. 5 provides the corresponding MAE results for Albumin, LDH, and Platelet Count, showing similar horizon-dependent degradation patterns. Error bars represent 95% confidence intervals.

C Dynamic rediction Performance Using Individual Laboratory Values in the OMU Dataset

Fig. 6 summarizes the predictive accuracy in AUC obtained for each type of longitudinal laboratory values when used individually as the sole predictor in the OMU dataset. Here, we focus particularly on the CRP and LDH results. In both CRP and LDH results, the yellow bars (Mean+Std) substantially exceed the red baseline line, indicating that combining the mean and the standard deviation markedly improves predictive performance. Clinicians note that CRP and LDH often exhibit threshold-like behavior, where the mortality risk increases sharply once the laboratory value exceeds a certain level. Such abrupt changes are difficult to capture using the mean alone, whereas the standard deviation can reflect variability or transient spikes in the measurements. As a result, for these laboratory variables, the standard deviation likely carries additional prognostic information, which may explain the performance gains observed in the “Mean+Std” model.

D Proposed method vs. Landmarking 2.0

As defined in the introduction, L denotes the landmark time at which a prediction is issued, and $\{X_t\}_{t\geq 0}$ denotes the longitudinal laboratory measurement process. The laboratory history observed up to the landmark is written as $\mathcal{X}_L = \{X_s : s \leq L\}$. Landmarking 2.0 fits a working longitudinal model to the observed measurements $\mathcal{X}_L := \{X_s :\ s \leq L\}$ and derives a patient-specific predicted future trajectory $\hat{X}(u \mid L)$ for $u > L$. The time-dependent Cox model uses $\hat{X}(u \mid L)$ as the covariate value at each time u.

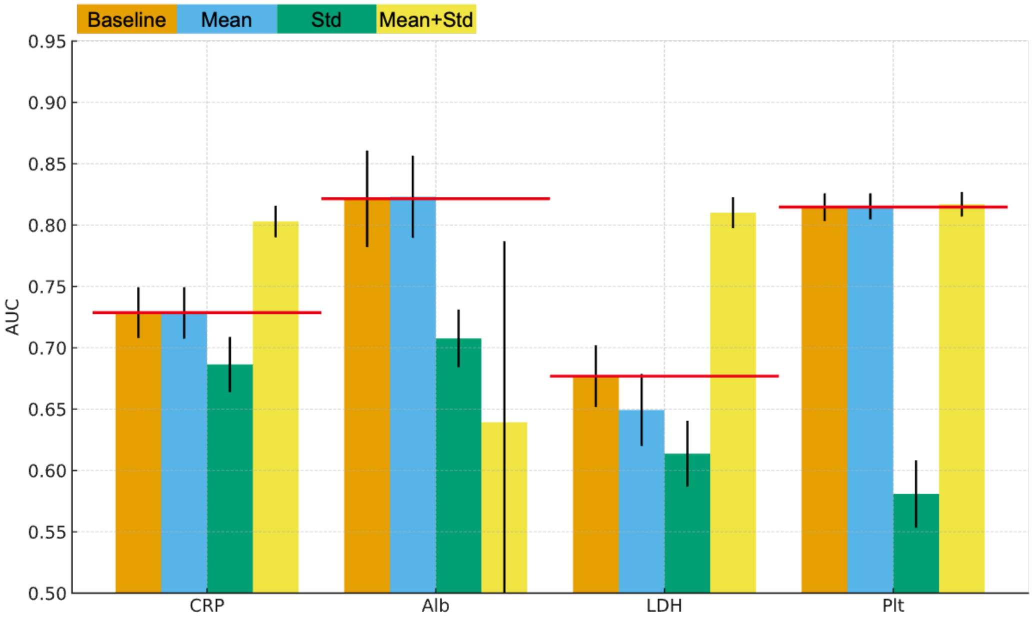


Figure 6: This figure compares predictive accuracy obtained on the OMU dataset when each type of longitudinal laboratory values is used individually as the sole predictor. The orange bars represent the baseline method, the blue bars the predictive accuracy only with the mean, the green bars the predictive accuracy only with the standard deviation, and the yellow bars the predictive accuracy with the mean and the standard deviation. The red horizontal line indicates the baseline AUC for reference.

By contrast, Stage 1 of our method fits a single regression model across all patients to predict a near-future laboratory value X_{L+k} for each individual. The resulting forecast $\widehat{X}_{L+k}$ is then used in Stage 2 as an additional covariate when predicting the d-day outcome, with the explicit aim of shortening the effective prediction horizon from the landmark time.

In summary, Landmarking 2.0 uses a patient-specific predicted trajectory $\hat{X}(u \mid L)$ as a time-dependent covariate and models the full d-day horizon directly from L, whereas our method relies on a single near-future prediction $\widehat{X}_{L+k}$, treated as a fixed covariate, and focuses prediction on a shorter effective horizon. This contrast captures the essential difference between Landmarking 2.0 and our two-stage approach.