2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) | 978-1-6654-0358-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/BHI50953.2021.9508537

Real-time Mortality Prediction Using MIMIC-IV ICU Data Via Boosted Nonparametric Hazards

Zhale Nowroozilarki*, Arash Pakbin*, James Royalty*, Donald K.K. Lee†, and Bobak J. Mortazavi*
*Department of Computer Science & Engineering, Texas A&M University,
Email: {zhale, a.pakbin, troyalty18, bobakm}@tamu.edu
†Goizueta Business School and Dept of Biostatistics & Bioinformatics, Emory University, Email: donald.lee@emory.edu

***Abstract*—Electronic Health Record (EHR) systems provide critical, rich and valuable information at high frequency. One of the most exciting applications of EHR data is in developing a real-time mortality warning system with tools from survival analysis. However, most of the survival analysis methods used recently are based on (semi)parametric models using static covariates. These models do not take advantage of the information conveyed by the time-varying EHR data. In this work we present an application of a highly scalable survival analysis method, BoXHED 2.0 [1], to develop a real-time in-ICU mortality warning indicator based on the MIMIC IV data set [2]. Importantly, BoXHED can incorporate time-dependent covariates in a fully nonparametric manner and is backed by theory [3]. Our in-ICU mortality model achieves an AUC-PRC of 0.41 and AUC-ROC of 0.83 out of sample, demonstrating the benefit of real-time monitoring.**

***Index Terms*—Electronic Health Record, Survival analysis, Hazard estimation, Nonparametric, Time-dependent covariates, MIMIC IV Dataset**

I. Introduction

Electronic Health Records (EHRs) and other health information technology (e.g. personal data from wearable sensing) provide a potential trove of data for clinical risk modeling. Its real-time nature represent a major advantage over administrative or registry-based data [4]. However, the development of clinical models tends to remain within large registries that abstract the data into static snapshots of patient health [5]. Even machine learning models fail to substantively improve the prediction on these time-static datasets [6], highlighting the need for methods that leverage richness of the EHR data to improve performance [7].

With the availability of ICU data in the form of the MIMIC-III and MIMIC-IV datasets [8], models that take advantage of EHR data have gained prominence. While these models show promise in predicting important clinical outcomes such as mortality [9], [10], they tend to only generate one classification prediction at one point in time during the entire episode of care. For example, [11] uses the first 24 hours of data to predict outcomes after cardiovascular procedures. Prediction systems for in-ICU mortality that are more dynamic in nature update forecasts periodically using the latest information available [12], [13].

Ideally, adverse event warning systems should operate in real-time as a patient's episode of care evolves. An example of this can be seen in the prediction of sepsis in admitted patients [14]. However, existing real-time prediction methodologies are based on classical statistical models that do not take advantage of recent advances in machine learning. The purpose of this paper is to explore the performance improvement that can be gained from embedding state-of-the-art survival analysis techniques into real-time mortality warning systems.

We focus on a recent survival methodology called BoXHED [1], [15], which is a gradient boosted procedure that is well suited to estimating clinical risk in the presence of time-varying features from EHRs. We train BoXHED to the MIMIC-IV dataset, and use it to create an in-ICU mortality warning system that continually assesses risk as the features evolve. This is particularly relevant to patients with short stays (under 5 days) as they have the most variable conditions. Out-of-sample performances are compared to two benchmarks: The classic time-varying Cox model, and a method that is representative of recent deep learning approaches for survival data (Dynamic DeepHit).

II. Survival models for time-varying features

Techniques used to forecast the time T to an event (e.g. mortality) fall under the survival analysis discipline. When time-dependent features $X(t)$ are involved, it is shown in [15] that the fundamental quantity of interest is the hazard function $\lambda(t, x)$, which is the conditional probability of the event occurring in the next instant given that it has not yet occurred:

$$\lambda(t,x)dt \approx \mathbb{P}(T \in [t, t+dt) | T \geq t, X(t) = x). \quad (1)$$

Thus $\lambda(t, X(t))$ is the most natural measure of real-time mortality risk. Note that $X(t)$ can either be the current values of the features, or it can be feature-engineered to be its history up to t.

A. Cox proportional hazards model

The venerated Cox model [16] is the workhorse model used in applications, but it imposes a key assumption on the functional form of $\lambda(t, x)$ that rules out potential interaction effects between time and the covariates:

$$\lambda_{PH}(t,x) = h_0(t)e^{R(x)}. \quad (2)$$

The baseline hazard function $h_0(t)$ is difficult to estimate from data without further assumptions, but $R(x)$ can be estimated independently of $h_0(t)$. Thus $R(x)$ provides a relative risk

This work was supported in part by NIH grant 1R21EB028486-01.

978-1-6654-0358-0/21/$31.00 ©2021 IEEE
Authorized licensed use limited to: Texas A M University. Downloaded on July 15,2024 at 17:10:07 UTC from IEEE Xplore. Restrictions apply.

score that can be used to compare subjects, but an absolute risk score is not readily available.

Traditionally, $R(x) = \beta' x$ is modeled linearly and this specification permits the inclusion of time-varying features [17]. Following the sepsis prediction application [14], we use $R(X(t))$ as a real-time measure of (relative) mortality risk for the Cox model benchmark.

We note that recent works in machine learning relax the linear specification by using deep neural networks to model $R(x)$ [18], [19]. However, these methods only accommodate time-static covariates and are therefore unable to take advantage of the information conveyed by the time-varying clinical data.

B. Deep learning survival models

In a step towards accommodating time-dependent covariates in a nonparametric way, deep survival models forgo real-time prediction by specializing to discrete time τ [20], [21]. Survival prediction then becomes solving binary classification problems at each point on a predefined time grid. For example, Dynamic DeepHit [21] further imposes a predefined time $\tau_{\max}$ by which the event has to happen with probability one, so that a recurrent neural network with a soft-max output layer can be used to estimate

$$o_\tau := \mathbb{P}(T = \tau | \mathcal{X}) \tag{3}$$

for $\tau \leq \tau_{\max}$ and $\sum_{\tau \leq \tau_{\max}} o_\tau = 1$. Here, $\mathcal{X}$ is the covariate history up to the time of prediction. Extension to competing risks is also studied in [21].

To use Dynamic DeepHit to produce a mortality risk measure at time t for our deep learning benchmark, we discretize continuous time into hourly bins $\tau_1, \cdots, \tau_{\max}$ and denote the bin containing t as $\tau(t)$. The risk measure at t is then taken to be $o_{\tau(t)}$, which is the discrete approximation to the hazard (1) at t.

We note that the chief purpose of [21] is not to estimate (3), but to use them to estimate the cumulative probabilities $\mathbb{P}(T \leq \tau | \mathcal{X})$. For our application, the cumulative probability is not an appropriate risk measure because our prediction target is whether or not a patient dies in the ICU at *any* point after t. Since Dynamic DeepHit assumes that the event must occur by $\tau_{\max}$ hours in the ICU ($\sum_{\tau \leq \tau_{\max}} o_\tau = 1$), the predicted cumulative probability of in-ICU death will always be 1.

C. Boosted nonparametric hazards

Very recently, [3] developed a theoretically justified gradient boosting solution for estimating the hazard (1) nonparametrically with (continuous) time-varying features. A scalable tree-boosted implementation called BoXHED can handle recurrent events as well as survival data beyond right-censoring [1]. Support for missing data and multicore CPU/GPU computing are also included. BoXHED performs regularized minimization of the negative nonparametric likelihood, and does so by iteratively adding shallow regression trees. In contrast to the Cox model, the BoXHED hazard estimator $\hat{\lambda}(t, x)$ provides an absolute measure of a subject's real-time mortality risk rather than a relative one.

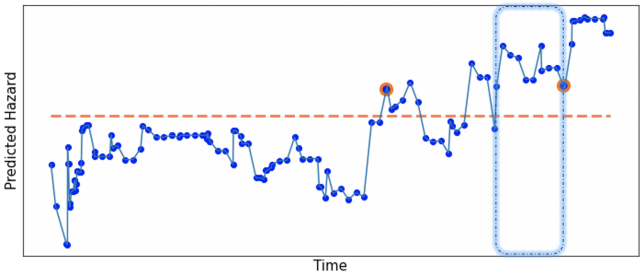

Fig. 1. Generating real-time mortality predictions from the values of a patient's risk measure (blue). The first orange dot marks the first time the risk measure exceeds the threshold (dashed horizontal line). The second orange dot marks the first time the risk measure stayed above the threshold for 8 hours (rectangle window).

III. Methods

We compare the performance of BoXHED to those of the baselines (time-varying Cox and Dynamic DeepHit) at predicting in-ICU mortality on a continuous basis. The data comes from MIMIC IV [8]. We follow the approach in the sepsis prediction application [14] to convert survival risk measures into real-time mortality predictions, which is to use the classic sliding window to update risks. While the Cox relative risk $R(X(t))$ was used in [14], this work evaluates the improvements brought specifically by BoXHED's boosted nonparametric hazards approach.

For BoXHED, real-time mortality predictions are generated from the values of the risk measure $\hat{\lambda}(t, X(t))$ over time in the following way (the same approach applies to the risk measures produced by the other two methods): For a given risk threshold ρ, we look at whether the patient's risk measure is above ρ for the past 8 hours. A patient is then flagged as predicted to eventually die in the ICU if this is true (as illustrated by the second orange dot in Figure 1). Once flagged, no further predictions are made for the patient in question. A second criterion is to flag a patient as soon as the risk measure exceeds ρ (as illustrated by the first orange dot in Figure 1). Further details are provided in the following subsections.

A. Data

Due to structural differences between the MIMIC III and MIMIC IV datasets, we extract the MIMIC IV dataset using a modified version of the preprocessing pipeline introduced in [9] for MIMIC III. The modified pipeline merges MIMIC IV's various tables in order to derive patient ICU history. This results in 31,544 ICU stays, of which two are removed as outliers as they have many more measurements compared to others. All told, this paper focuses on 31,542 ICU stays.

Furthermore, we focus on only the first 120 hours (5 days) of each ICU stay for two reasons. First, the required computational effort for Dynamic DeepHit explodes for a large number of discrete time periods. Second, early intervention is significantly associated with positive patient outcomes [22], which makes real-time monitoring particularly valuable during

Authorized licensed use limited to: Texas A M University. Downloaded on July 15,2024 at 17:10:07 UTC from IEEE Xplore. Restrictions apply.

the first few days. After the initial period, overall patient status is better understood.

The 17 predictive features used to fit our models come from [9]: Capillary refill rate, Diastolic blood pressure, Fraction inspired oxygen, Glascow coma scale eye opening, Glascow coma scale motor response, Glascow coma scale total, Glascow coma scale verbal response, Glucose, Heart rate, Height, Mean blood pressure, Oxygen saturation, Respiratory rate, Systolic blood pressure, Temperature, Weight and pH. Since these features can be found in many datasets, this makes it easy to compare our results to studies based on other datasets.

B. Training

The 31,542 ICU stays are randomly split into training and testing sets according to a 80/20 split of the unique patient IDs. This is because one patient may contribute to multiple ICU stay records, so we need to avoid assigning such a patient's second ICU stay to the training set and their first stay to the testing set. All three methods are fit to the training set.

For training BoXHED 2.0, we use the built-in K-fold cross-validation function (with $K = 5$) to select the number of trees and the depth of the trees to use. Using the one-standard-error rule (§7.10 of [23]) to select the most parsimonious model within one standard error of the best performing one, we arrive at using 75 trees of maximum depth 2 (i.e. 4 leaf nodes max).

For training the time-varying Cox model, we use the Lifelines package [24] in Python. Unlike BoXHED 2.0 and Dynamic DeepHit, this package does not automatically handle missing data. We therefore impute missing values in the same way as [9]: If a previous measurement exists, its value is carried forward. Otherwise, the missing feature is imputed using a pre-defined value.

C. Scoring the predictions

As explained earlier, the thresholding criterion used to flag in-ICU mortality for stays in the testing set is continuously assessed as new data stream in. If the flag is raised during the first 120 hours of the stay, a positive prediction is made for the stay. Otherwise, a negative prediction is made. This approach converts the time-varying output from a dynamic survival model (i.e. the risk measure) into a classification signal. To compute the area under the receiver operating characteristic curve (AUC-ROC) and the area under the precision recall curve (AUC-PRC) for the testing set predictions, the threshold ρ is varied to trace out both curves.

IV. Results

Table I presents the out-of-sample performances for mortality predictions triggered by the risk measure exceeding the threshold at any point in time. As a reminder, the AUC-ROC baseline of 0.50 corresponds to a random guess, and the AUC-PRC baseline of 0.09 corresponds to always predicting positive. While AUC-ROC is commonly used to evaluate classifiers, AUC-PRC is more informative here (and often in clinical datasets) given the imbalance between the number of negative and positive outcomes [25].[1]

TABLE I
COMPARISON OF MODEL PERFORMANCES: PREDICTIONS BASED ON RISK MEASURE EXCEEDING THRESHOLD AT ANY TIME (*SEE § IV FOR DISCUSSION OF DYNAMIC DEEPHIT RESULTS)

Model	AUC-ROC	AUC-PRC
Baseline	0.50	0.09
Time-varying Cox	0.74	0.29
Dynamic DeepHit*	0.50	0.06
BoXHED	**0.78**	**0.35**

TABLE II
COMPARISON OF MODEL PERFORMANCES: PREDICTIONS BASED ON RISK MEASURE EXCEEDING THRESHOLD FOR 8 HOURS (*SEE § IV FOR DISCUSSION OF DYNAMIC DEEPHIT RESULTS)

Model	Window Size	AUC-ROC	AUC-PRC
Baseline	-	0.50	0.09
Time-varying Cox	8hrs	0.81	0.36
Dynamic DeepHit*	8hrs	0.47	0.05
BoXHED	8hrs	**0.83**	**0.41**

We see from Table I that BoXHED handily outperforms both Time-varying Cox and Dynamic DeepHit, particularly on AUC-PRC. However, a caveat is required for the seemingly dismal performance of Dynamic DeepHit. This could be due to the fact that the method assumes that mortality must occur by some time $\tau_{\max}$ in the ICU, which is inherently incompatible with the current application since 91% of stays do not end in death. Another possible explanation is suboptimal hyperparameter tuning, which is not as systematic for deep learning as it is for BoXHED. Indeed, tuning neural nets is an art as there are far more degrees of freedom, all the way up to modifying the network architecture itself.

Table II presents the out-of-sample performances for predictions based on having the risk measure remain above the threshold for 8 hours. Figure 2 illustrates the precision-recall curve for this case. The direction of the results are qualitatively the same as those in Table I, except that Time-varying Cox and BoXHED's performances are noticeably better. This is intuitive, since requiring the risk measure to remain elevated for a longer period reduces the number of false positives.

V. Limitations and Future Work

Rather than focus on the most recent 8 hours of risk measure values, a moving average might capture more information. Future work could explore the potential of flagging patients if the moving average exceeds some threshold. Second, the current prediction target is whether or not a patient will eventually die in the ICU, be it in 2 hours or 2 days. To help physicians prioritize care for patients at higher imminent risk, future research should consider shorter prediction horizons such as in-ICU death within 6 hours [12]. Lastly, the outcome

[1]The benchmark classical prediction model for the decompensation task in MIMIC, the closest to our use case, achieves an AUC-PRC of 0.34 [9].

Authorized licensed use limited to: Texas A M University. Downloaded on July 15,2024 at 17:10:07 UTC from IEEE Xplore. Restrictions apply.

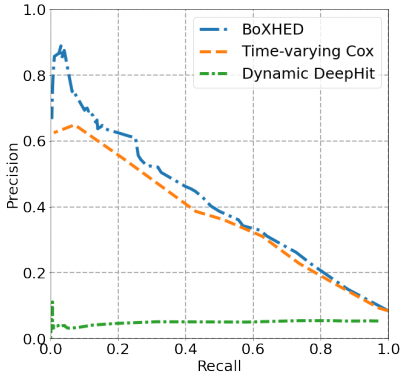

Fig. 2. Precision-Recall curve for risk exceeding threshold for entirety of past 8 hours

of interest should be expanded from in-ICU mortality to in-hospital mortality, and to account for competing risks from multiple adverse events, potentially over different time scales.

VI. Conclusion

EHR data is a rich source of information for adverse event prediction in clinical settings. The high-frequency, time-varying data present opportunity to develop real-time warning systems that update estimates of patient mortality hazards with the introduction of each new data point. Survival analysis is the ideal tool for this. However, there is a dearth of survival methods that can handle time-varying features non-parametrically and at scale. This work presents the application of such a tool called BoXHED for developing an in-ICU mortality warning system using MIMIC-IV data. The system achieves state-of-the-art results (AUC-ROC 0.83, AUC-PRC 0.41) when compared to the benchmarks. The results highlight the promise of BoXHED, a gradient-boosted nonparametric hazard estimator, for real-time clinical predictions.

References

[1] A. Pakbin, X. Wang, B. J. Mortazavi, and D. K. K. Lee, "BoXHED 2.0: Scalable boosting of dynamic survival analysis," *arXiv preprint arXiv:2103.12591*, 2021.

[2] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[3] D. K. K. Lee, N. Chen, and H. Ishwaran, "Boosted nonparametric hazards with time-dependent covariates," *Annals of Statistics (forthcoming)*.

[4] W. L. Schulz, J. C. Kvedar, and H. M. Krumholz, "Agile analytics to support rapid knowledge pipelines," 2020.

[5] R. L. McNamara, K. F. Kennedy, D. J. Cohen, D. B. Diercks, M. Moscucci, S. Ramee, T. Y. Wang, T. Connolly, and J. A. Spertus, "Predicting in-hospital mortality in patients with acute myocardial infarction," *Journal of the American College of Cardiology*, vol. 68, no. 6, pp. 626–635, 2016.

[6] R. Khera, J. Haimovich, N. C. Hurley, R. McNamara, J. A. Spertus, N. Desai, J. S. Rumsfeld, F. A. Masoudi, C. Huang, S.-L. Normand *et al.*, "Use of machine learning models to predict death after acute myocardial infarction," *JAMA cardiology*.

[7] M. M. Engelhard, A. M. Navar, and M. J. Pencina, "Incremental benefits of machine learning—when do we need a better mousetrap?" *JAMA cardiology*, 2021.

[8] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. Celi, and R. Mark, "MIMIC-IV (version 0.4)," 2020.

[9] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[10] A. Pakbin, P. Rafi, N. Hurley, W. Schulz, M. H. Krumholz, and J. B. Mortazavi, "Prediction of icu readmissions using data at patient discharge," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 4932–4935.

[11] B. J. Mortazavi, N. Desai, J. Zhang, A. Coppi, F. Warner, H. M. Krumholz, and S. Negahban, "Prediction of adverse events in patients undergoing major cardiovascular procedures," *IEEE journal of biomedical and health informatics*, vol. 21, no. 6, pp. 1719–1729, 2017.

[12] J. Ma, D. K. K. Lee, M. E. Perkins, M. A. Pisani, and E. Pinker, "Using the shapes of clinical data trajectories to predict mortality in ICUs," *Critical care explorations*, vol. 1, no. 4, 2019.

[13] S. Wang, M. B. McDermott, G. Chauhan, M. Ghassemi, M. C. Hughes, and T. Naumann, "MIMIC-extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 222–235.

[14] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (TREWScore) for septic shock," *Science translational medicine*, vol. 7, no. 299, pp. 299ra122–299ra122, 2015.

[15] X. Wang, A. Pakbin, B. Mortazavi, H. Zhao, and D. K. K. Lee, "BoXHED: Boosted eXact Hazard Estimator with Dynamic covariates," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9973–9982.

[16] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[17] Z. Zhang, J. Reinikainen, K. A. Adeleke, M. E. Pieterse, and C. G. Groothuis-Oudshoorn, "Time-varying covariates and coefficients in cox regression models," *Annals of translational medicine*, vol. 6, no. 7, 2018.

[18] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, pp. 1–12, 2018.

[19] C. Nagpal, S. Yadlowsky, N. Rostamzadeh, and K. Heller, "Deep cox mixtures for survival regression," *arXiv preprint arXiv:2101.06536*, 2021.

[20] K. Ren, J. Qin, L. Zheng, Z. Yang, W. Zhang, L. Qiu, and Y. Yu, "Deep recurrent survival analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4798–4805.

[21] C. Lee, J. Yoon, and M. Van Der Schaar, "Dynamic-DeepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 122–133, 2019.

[22] J.-U. Song, G. Y. Suh, H. Y. Park, S. Y. Lim, S. G. Han, Y. R. Kang, O. J. Kwon, S. Woo, and K. Jeon, "Early intervention on the outcomes in critically ill cancer patients admitted to intensive care units," *Intensive care medicine*, vol. 38, no. 9, pp. 1505–1513, 2012.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[24] C. Davidson-Pilon, J. Kalderstam, N. Jacobson, S. Reed, B. Kuhn, P. Zivich, M. Williamson, AbdealiJK, D. Datta, A. Fiore-Gartland, A. Parij, D. WIlson, Gabriel, L. Moneda, A. Moncada-Torres, K. Stark, H. Gadgil, Jona, K. Singaravelan, L. Besson, M. S. Peña, S. Anton, A. Klintberg, GrowthJeff, J. Noorbakhsh, M. Begun, R. Kumar, S. Hussey, S. Seabold, and D. Golland, "Camdavidsonpilon/lifelines: v0.25.11," Apr. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4683730

[25] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

Authorized licensed use limited to: Texas A M University. Downloaded on July 15,2024 at 17:10:07 UTC from IEEE Xplore. Restrictions apply.