Robust Real-Time Mortality Prediction in the Intensive Care Unit using Temporal Difference Learning

Thomas Frost THOMAS.FROST.21@UCL.AC.UK
Ken Li KEN.LI@UCL.AC.UK
Steve Harris STEVE.HARRIS@UCL.AC.UK
Institute of Health Informatics, University College London, United Kingdom

Abstract

The task of predicting long-term patient outcomes using supervised machine learning is a challenging one, in part because of the high variance of each patient's trajectory, which can result in the model over-fitting to the training data. Temporal difference (TD) learning, a common reinforcement learning technique, may reduce variance by generalising learning to the pattern of *state transitions* rather than terminal outcomes. However, in healthcare this method requires several strong assumptions about patient states, and there appears to be limited literature evaluating the performance of TD learning against traditional supervised learning methods for long-term health outcome prediction tasks. In this study, we define a framework for applying TD learning to real-time irregularly sampled time series data using a Semi-Markov Reward Process. We evaluate the model framework in predicting intensive care mortality and show that TD learning under this framework can result in improved model robustness compared to standard supervised learning methods – and that this robustness is maintained even when validated on external datasets. This approach may offer a more reliable method when learning to predict patient outcomes using high-variance irregular time series data.

Keywords: Predictive Models, Deep Learning, Reinforcement Learning, Intensive Care, Time Series.

Data and Code Availability This research makes use of the MIMIC-IV dataset (Johnson et al., 2020, 2023) and the Salzburg Intensive Care dataset (Rodemund et al., 2023a,b), two datasets of de-identified health data collected from Beth Israel Deaconess Medical Center (USA) and University Hospital Salzburg (Austria), respectively. Both datasets are available via the PhysioNet platform (Goldberger et al., 2000). Code for this paper is publicly available at https://github.com/tdgfrost/td-icu-mortality.

Institutional Review Board (IRB) For both datasets, the collection of patient information and sharing of de-identified health data for the purposes of research received local ethical approval, as detailed in their respective publications. The research contained herein did not require further independent ethical approval.

1. Introduction

Patients in the Intensive Care Unit (ICU) are amongst the sickest in any hospital and often follow a complex trajectory from admission to discharge (or death). With an average ICU mortality rate between 7-19%, there is a need for accurate prediction models to help stratify patient risk during admission (Checkley et al., 2014; Capuzzo et al., 2014).

Traditional approaches rely heavily on scoring systems such as APACHE, SOFA, and SAPS, which provide a mortality risk score from patient features collected at ICU admission. These methods have been externally validated but are cited to have area under the receiver operating characteristic curve (AUROC) scores frequently limited to the 0.70-0.79 range (Ko et al., 2018; Mandrekar, 2010; Sarkar et al., 2021).

Supervised machine learning algorithms (such as gradient boosted ensembles, artificial neural networks, and Bayesian networks) demonstrate improvement on this baseline, with AUROC scores ranging from 0.80-0.95. However, many of these studies lack high-quality external validation; are limited to one-time predictions at the point of admission (as opposed to ongoing/real-time

predictions); and/or limit their mortality time-horizon to the short-term (≤72 hours) (García-Gallo et al., 2020; Nistal-Nuño, 2022; Iwase et al., 2022; Lei et al., 2023; Kim et al., 2019). Research studies that have validated their models externally typically report a significant deterioration in AUROC performance (10-15%), consistent with over-fitting to the training data (Lei et al., 2023; Meyer et al., 2018).

We theorise that the over-fitting of models in these tasks may be related to the significant variance of each patient's trajectory, with diminishing relevance of current features to outcomes far in the future. In reinforcement learning (the branch of machine learning tasked with optimal sequential decision-making), models face a similar challenge of attributing distant rewards under noisy trajectories to present states and actions (the so-called "credit assignment problem"). A common solution to this has been the use of temporal difference (TD) learning, in which the model is bootstrapped using its own predictions for future states rather than the actual observed (distant) rewards (Sutton, 2018). This is demonstrated conceptually in Figure 1.

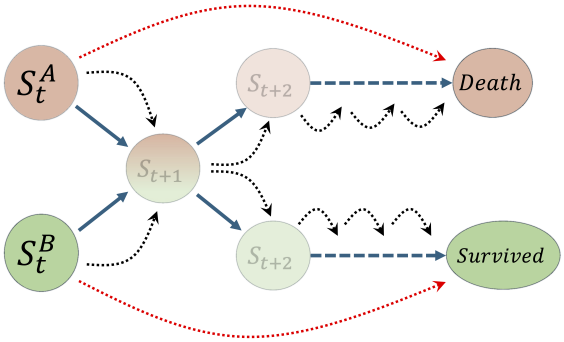

Figure 1: Illustrative example of TD learning versus supervised learning. Imagine two patient trajectories (A and B), which share a similar state at S_{t+1}. A model trained using supervised learning might learn to predict a mortality risk of 100% for S_t^A, and 0% for S_t^B, based on the observed real outcomes for each state (long red dotted line). A model trained using TD learning would instead learn a 50% risk for both states, as each inherits the aggregate predicted risk of all future states (short black dotted lines).

Several reinforcement learning publications already exist exploring the use of TD learning in predicting mortality risk for a given set of actions in intensive care (Komorowski et al., 2018; Prasad et al., 2017; Liu et al., 2024; Wang et al., 2022b). However, to date these works focus on using TD learning to generate *counterfactual* predictions based on a *hypothetical* decision policy, making it difficult to evaluate their accuracy without real-world policy deployment (Levine et al., 2020). Additionally, most such studies artificially aggregate time series data into regular intervals (e.g., every four hours), with only one exception (Kim et al., 2021) applying TD learning to irregular health data.

In this paper, we generate a set of models using either TD learning or supervised learning for the task of inpatient mortality prediction. The models are trained on more than 65,000 undifferentiated patients using the MIMIC-IV dataset. We describe the mathematical framework for patient states and state transitions (Section 2.3 and 2.1), the CNN-LSTM base model architecture (Section 2.4.1), and the training process for each model category (Section 2.4.3). We report the AUROC scores for all models on both a hold-out test segment of the MIMIC-IV dataset, as well as an external validation dataset (SICdb) of 21,000 patients (Section 3.3 and 3.4). We show that models trained with TD learning outperform both supervised learning and clinical score baselines for the task of long-term mortality prediction, and suffer much less overfitting when tested on an external validation dataset. We discuss our interpretations of these results and plans for further work in this area (Section 4).

2. Methods

2.1. Temporal Difference Learning and Semi-MRP Framework

2.1.1. Markov Reward Process

The Markov Reward Process (MRP) describes a Markov chain of states, in which "memoryless" states transition stochastically to new states over time, and generate rewards $\mathcal{R}$ in the process (similar to Figure 1). It is defined by the tuple $\{\mathcal{S}, \mathcal{P}_{ss'}, \mathcal{R}, \gamma\}$:

- $\mathcal{S}$: The current state.
- $\mathcal{P}_{ss'}$: The state transition probability matrix from $\mathcal{S}$ to $\mathcal{S}'$, which is independent of any states prior to $\mathcal{S}$.

- $\mathcal{R}$: The immediate reward received.
- γ: A time discount factor, $\lambda \in [0, 1]$.

In healthcare, the latent state $\mathcal{S}_t$ of the patient is not observable, but could theoretically be inferred by the model from observations $\mathcal{O}_t$ recorded over a suitable time period. The transition matrix $\mathcal{P}_{ss'}$ is thus also unknown, but is expected to be affected by the decision policy of the unobserved clinician.

2.1.2. TD Learning

One can train a value function $V(O_t)$ to predict the expected sum of (discounted) rewards $\mathcal{R}$ collected at all future steps from a set of observations $\mathcal{O}_t$ as follows (1):

$$V(\mathcal{O}_t) \leftarrow \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}_{t+k+1}\right] \tag{1}$$

This equation requires complete episode trajectories using Monte Carlo sampling, which is unbiased but potentially of high variance, particularly as the future trajectory becomes very long. In contrast, via mathematical induction of (1), we arrive at (2), the formula for TD learning. This equation uses bootstrapping to reduce the variance by predicting a single step in the future (at the risk of bias).

$$V(\mathcal{O}_t) \leftarrow \mathbb{E}\left[\mathcal{R}_{t+1} + \gamma V(\mathcal{O}_{t+1})\right] \tag{2}$$

2.1.3. Semi-Markov Reward Process

One limit of MRPs is their dependence on state transitions modelled under regular discrete intervals. To allow for irregularly sampled health data, we instead use a semi-MRP (3), in which transitions are allowed to occur at variable intervals:

$$V(\mathcal{O}_t) \leftarrow \mathbb{E}_{k\sim\mathcal{D}}\left[\bar{\mathcal{R}}_{t:t+k} + \gamma^k V(\mathcal{O}_{t+k})\right] \tag{3}$$

We assume that k is sampled from a (potentially unknown) stochastic distribution $\mathcal{D}$, with $\bar{\mathcal{R}}_{t:t+k}$ representing the discounted sum of rewards captured in the interval between t and $t + k$.

2.1.4. Choice of Reward under SMRP

We can now convert the formula in Equation (3) to a loss function for predicting the terminal outcome (death/discharge) of a patient. We define the following:

- The terminal reward is patient mortality, with $\mathcal{R} = 1$ for death and $\mathcal{R} = 0$ for survival (i.e., successful discharge).
- There are no interim rewards prior to the terminal state.
- $\gamma = 1$, i.e., the predicted risk of death is the averaged risk from all immediate future states.

The above conditions allow us to convert (3) to (4), producing an averaged mortality risk at time t from 0 to 100%:

$$V(\mathcal{O}_t) \leftarrow \mathbb{E}_{k\sim D}\left[\begin{cases}\mathcal{R}, & \text{if } \mathcal{O}_t \text{ is terminal}\\ V(\mathcal{O}_{t+k}), & \text{otherwise}\end{cases}\right] \tag{4}$$

We will now describe the processing steps to accommodate health data into this framework.

2.2. Data Selection and Pre-Processing

2.2.1. MIMIC-IV

The Medical Information Mart for Intensive Care IV (MIMIC-IV v3.0) database is a collection of de-identified electronic healthcare data for more than 360,000 patients admitted to Beth Israel Deaconess Medical Center, USA between 2008 - 2022 (Johnson et al., 2020, 2023). Our training data consisted of 65,000 patients across 85,000 hospital admissions with at least one of the required input features. These included a mixture of elective post-operative and emergency surgical/medical admissions. The input features consisted of a range of biomarkers, intravenous medications, and demographic information (Appendix B). The patients were divided into train (80%), validation (10%), and test (10%) groups. All input data were standardised based on mean and standard deviation values computed from the training group. Additional per-feature processing steps are described in Appendix C.

2.2.2. SICDB

For external validation, we employed the Salzburg Intensive Care database (SICdb v1.0.6), a publicly available European dataset containing de-identified healthcare data for more than 27,000 ICU admissions across 21,000 patients admitted to the University Hospital Salzburg between 2013 and

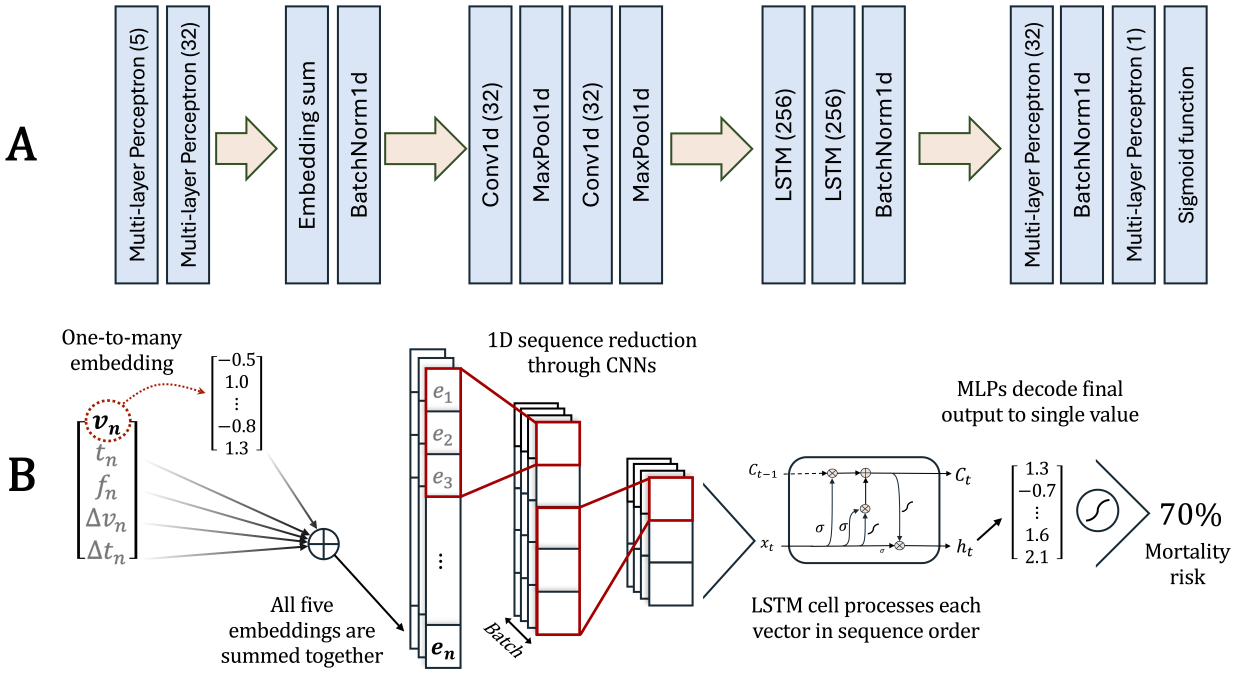

Figure 2: A) CNN-LSTM Model Architecture. B) Visualisation of how data flows through the model.

2021 (Rodemund et al., 2023a,b). The data were standardised using the mean and standard deviation values computed from the MIMIC-IV training group.

2.3. Data Representation and State Definition

2.3.1. Time Series as 1D Sequences

Each patient's time series data was represented as a 1D sequence of tuples $\{v, t, f, \Delta v, \Delta t\}$. Here, v represents the measurement value, t the time-point relative to the current state marker (2.3.2), and f the feature label. Additionally, Δv denotes the change in value (if available) since the last measurement of the same feature, and Δt indicates the time interval between these measurements.

2.3.2. State Markers and Input Data Construction

In our framework, each measurement in a patient's admission is treated as a unique "state marker", from which the model can learn the patient's latent state at that point in time. Specifically, the model is provided with the observation data $\mathcal{O}_t$, which consists of the state marker and up to 396 retrospective measurements taken over the previous 7 days (plus age, gender, and weight). The model processes the data as an ordered sequence of real measurements (2.3.1 and Figure 2B), with the temporal component encoded within the tuple rather than the sequence position itself. The chosen model architecture (2.4.1) can accommodate sequences of variable lengths, and thus there is almost no missing data or requirement for data imputation (with the exception of patient weight, described in Appendix C).

2.3.3. Choice of Next State

When determining our distribution $\mathcal{D}$, it is important that our step sizes are large enough to demonstrate a clear trajectory, but not so large as to suffer excessive variance. We defined an eligibility window of 24 hours, with the start of the window delayed x hours into the future. The "next state" marker is chosen as the first available measurement occurring inside this window - if no measurements are made within the eligible period, we assume the patient has reached a terminal state (i.e., the current state marker was measured in the final 24 hours before discharge or death). We experimented with a range of possible delays when training the TD model, with x

Descriptor	MIMIC-IV	SICdb
Unique patients	65,050	21,447
Unique hospital admissions	84,659	*not available*
Unique ICU admissions	93,508	27,213
ICU length-of-stay, mean days (std)	3.7 (5.4)	3.5 (6.4)
Hospital mortality* (%)	11,126 (11.9%)	2,127 (7.8%)
1-year mortality (%)	23,655 (27.9%)	5,071 (18.6%)
Median age (IQR)	66 (55 - 77)	70 (60 - 75)
Female (%)	28,496 (43.8%)	8,170 (38.1%)

**Relative to each ICU admission*

Table 1: Baseline characteristics of the datasets.

set as either 4, 16, 24, 48, 72, or 120 hours. The best-performing model (on the validation dataset) was then chosen for further analysis.

2.4. Models

2.4.1. Model Architecture

The model architecture is demonstrated in Figure 2A. The input data consists of a sequence of observed measurements, in the form of a measurement tuple $\{v, t, f, \Delta v, \Delta t\}$ (2.3.1). Each tuple is embedded using five one-to-many multi-layer perceptron (MLP) networks, corresponding to each component of the tuple, before summing the embeddings together to create an embedded measurement e_n (Tipirneni and Reddy, 2022). The sequence of embedded measurements is then processed through CNN (for sequence length reduction) and LSTM (for sequence processing) layers, a popular machine learning architecture which has the advantage of being able to process sequences of variable length. The final hidden state of the terminal LSTM is decoded by two densely connected MLPs to give a single output for mortality risk. ReLU activation is used in all hidden layers, with sigmoid activation for the final output layer.

2.4.2. Candidate Models

We trained six groups of models using the same base architecture but different learning targets: our TD model, and five baseline models trained with supervised learning. The TD model was trained according to Equation (4), with the terminal reward set to 28-day mortality in the terminal state. The target for the supervised models was each state's observed mortality at one of several pre-defined time horizons (1 day, 3 days, 7 days, 14 days, and 28 days). This allows for a comparative evaluation of TD model performance across a spectrum of well-defined temporal horizons.

2.4.3. Model Training

Each candidate model was independently trained and evaluated five times. We used a binary cross-entropy loss, but included an optional class balancing factor for the supervised baselines (in which the loss is weighted according to the normalised inverse class frequency). This can often optimise supervised learning performance when training on class-imbalanced datasets such as MIMIC-IV.

Using the same network for both the predicted and target values can lead to training instability. To address this, we implement a separate identical "target network" that provides stable targets for the main network, and is gradually synchronised with the main network after each update (Lillicrap, 2015). More details for this and other training hyperparameters can be found in Appendix D.

2.4.4. Model Evaluation

The discriminative ability of all models was evaluated on the internal and external test data using the area under the receiver operating characteristic curve (AUROC) for each of the possible mortality labels – i.e., how well does the model's predicted score discriminate between classes when the label is set to mortality at 1 day, 3 days, 7 days, 14 days, or 28 days. AUROC scores were also calculated using

the Sequential Organ Failure Assessment (SOFA) score for samples in MIMIC-IV to provide a clinical baseline.

2.5. Statistical Testing

The mean performance of the TD model was compared to the mean performance of each baseline model in each evaluation using a one-tailed paired Student's t-test, with Benjamini-Yekutieli correction to account for multiple testing under dependency (Benjamini and Yekutieli, 2001). No comparison was made between the baseline models themselves.

2.6. Software

All training with performed using Python 3.11. Models were custom-built using PyTorch 2.3.1, and evaluated using the TorchEval 0.0.7 evaluation and SciPy 1.13 statistical testing packages.

3. Results

3.1. Dataset Baseline Characteristics

The baseline characteristics of the two datasets are summarised in Table 1. Both datasets exhibit similar ICU lengths of stay, with comparable means and standard deviations in the number of days in ICU. The patient cohorts in both datasets share similar median ages and gender distributions, which suggests a level of comparability in terms of the general patient population.

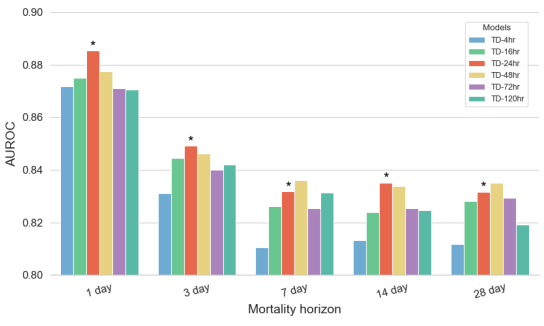

Figure 3: The effect of the size of interval between states during training on TD model performance (on the validation dataset). The chosen model, TD-24hr, is marked with an asterisk.

However, there is a notable difference in mortality outcomes, with the MIMIC-IV dataset showing a significantly higher average hospital and 1-year mortality rate. This may relate to differences in underlying co-morbidities and admission diagnoses for the two populations, as well as differences in the collection period.

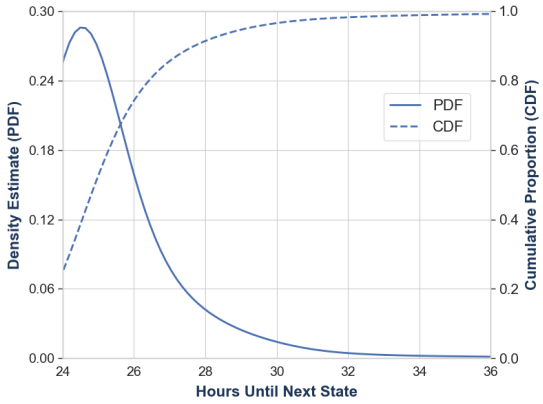

Figure 4: The distribution of time between "states" and "next states".

3.2. Effect of State Interval Size on TD Performance

The impact on TD model performance when trained with different sized delays between states is shown in Figure 3. The overall best-performing model was trained with a state-to-state delay of 24 hours (although a longer delay of 48 hours performed better for 7-day and 28-day mortality prediction). We thus chose the TD-24hr model as our benchmark model for subsequent analyses. For a 24-hour delay between states, the distribution of exact intervals between "states" and "next states" can be seen in Figure 4, with 50% of "next state" markers occurring within 1 hour of the eligible period, and 90% occurring within 4 hours.

3.3. Model Evaluation on the Internal Dataset

The evaluation results for each model group[1] on the MIMIC-IV dataset are summarised in Table

1. In nearly all cases, the supervised models trained with class weighting outperformed those without - full results are reported in Appendix A.

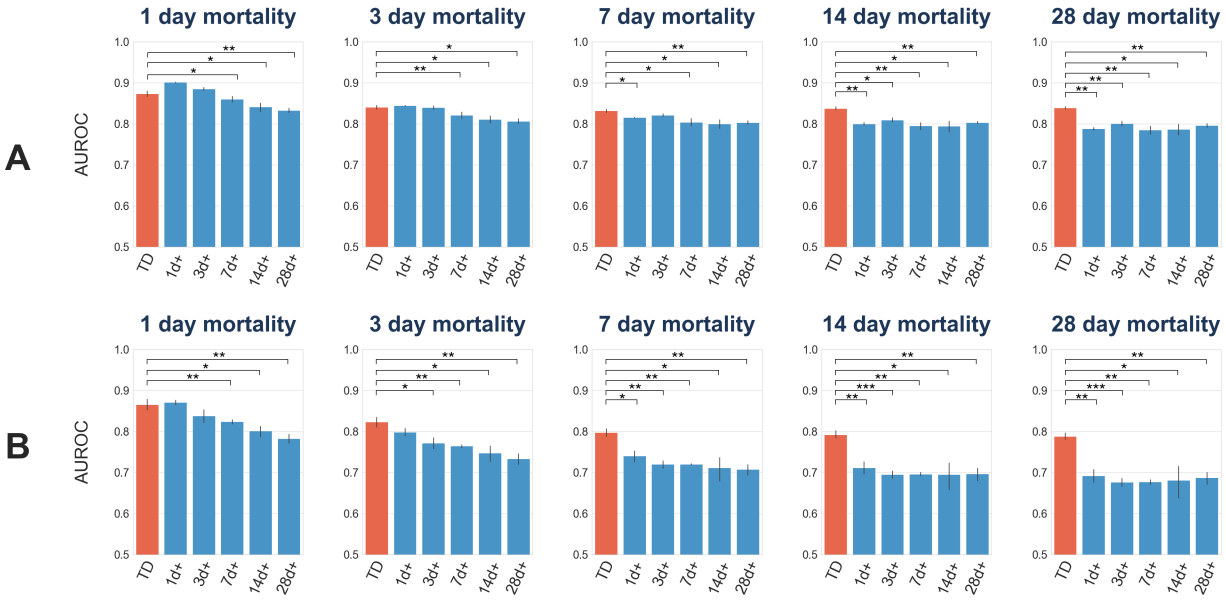

Figure 5: A) AUROC evaluation scores for the TD and supervised models$^+$ on different mortality horizons on the internal MIMIC-IV dataset. B) Evaluation scores for the same models on the external SICdb dataset.

$^*p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$
$^+$*only those trained with balanced cross-entropy shown.*

2 and Figure 5A. When predicting short-term mortality, the models trained via supervised learning on short-term mortality labels achieved the highest performance, with AUROC scores as high as 0.90. The predictions of the short-term models also had high AUROC scores for longer-term mortality, frequently outperforming the supervised models trained specifically on long-term labels. However, the TD model was found to consistently exceed the results of *all* baseline models when evaluated on longer time horizons (7 to 28 days). All models significantly outperformed the clinical SOFA score (0.69-0.74), which had AUROC scores consistent with other publications (Zhou et al., 2024; Esmaeili Tarki et al., 2023; Wang et al., 2022a).

3.4. External Model Validation

Each model was evaluated externally on the SICdb dataset, with results presented in Table 3 and Figure 5B. All models exhibited some degree of performance degradation on the unseen data, and this decline is most pronounced when predicting long-term mortality. When compared to the supervised models, the TD model showed noticeably less deterioration in predictive performance over the extended time horizons, and consistently outperformed all supervised models in all mortality predictions beyond the ultra-short (1-day) horizon.

		Models					
Mortality	SOFA	$1d^+$	$3d^+$	$7d^+$	$14d^+$	$28d^+$	*TD*
1-day	0.746	**0.901**	0.885	0.860	0.842	0.833	0.874
3-day	0.718	**0.844**	0.840	0.821	0.811	0.807	0.841
7-day	0.700	0.816	0.821	0.804	0.800	0.803	**0.832**
14-day	0.697	0.800	0.809	0.795	0.795	0.803	**0.838**
28-day	0.694	0.788	0.801	0.785	0.786	0.796	**0.839**

Table 2: Mean AUROC results on the internal MIMIC-IV dataset.
$^+$*indicates training with balanced cross-entropy.*

3.5. Statistical Analysis

Each model was trained five times, and the resulting mean and standard deviation was used to construct

	Models					
Mortality	$1d^+$	$3d^+$	$7d^+$	$14d^+$	$28d^+$	TD
1 day	**0.871**	0.838	0.824	0.801	0.783	0.865
3 days	0.799	0.772	0.765	0.748	0.734	**0.823**
7 days	0.741	0.720	0.720	0.712	0.708	**0.797**
14 days	0.712	0.695	0.696	0.695	0.697	**0.792**
28 days	0.692	0.676	0.677	0.681	0.688	**0.788**

Table 3: Mean AUROC results on the external SICdb dataset.
$^+$ *indicates training with balanced cross-entropy.*

95% confidence intervals, which are demonstrated in Figure 5. Most models were found to have relatively tight 95% CI bounds for their internal performance, with these bounds typically widening when evaluated on the external dataset. Student's t-test values were also calculated, with adjustments to account for multiple testing. The TD model outperformed the baseline supervised models to high degrees of statistical significance ($p \leq 0.01$), with the significance of these results increasing for longer prediction horizons and during external evaluation.

4. Discussion

In this report, we have demonstrated that models trained with temporal difference learning are able to produce long-term mortality predictions superior to those derived from conventional supervised learning (and far superior to clinical scores such as SOFA). This makes intuitive sense - as mentioned earlier, ICU patients can have lengths of stay that range from a matter of hours to many months, and an admission with a long duration can be expected to have greater variance for its final outcome. As machine learning algorithms become more complex, these models are then at risk of over-fitting to noise for these distant labels, and may deteriorate significantly when validated externally. We observe this in our results, with the supervised models trained directly on longer-term labels tending to perform worse on both the internal and external evaluations.

On the other hand, models trained on identical input data using temporal difference learning can learn more accurate predictions for long-term outcomes by generalising to the near-term trajectories of each state. This shortens the horizon of the target label to just a single step in the future, reducing the variance and degree of over-fitting, and does not appear to suffer significant bias for a sufficiently sized training dataset. Importantly, the prediction accuracy remains robust even when validated on an external dataset collected from another continent. When combined with our choice of model architecture, this produces a model that generalises well; can be applied at any stage in an ICU patient's admission; can process any length of inpatient stay; and does not suffer from data missingness.

From a purely theoretical standpoint, there are several reasons why TD learning could fail when applied to healthcare settings. First, the states are only partially observable, and it may not be possible for the model to infer an accurate latent state $\mathcal{S}_t$ for the patient given a limited set of visible observations $\mathcal{O}_t$. Second, the irregular distribution of time intervals (Figure 4) could limit the model's ability to consistently infer transitions between states. Third, two patients will never truly occupy the same continuous state space, and the model may fail to group "similar" patient observations together as part of trajectory mixing. Fourth, patient states may not observe a consistent transition probability over time, a key assumption of the SMRP. Despite all of these potential limitations, we have managed to show experimentally that models trained using TD learning are still able to converge to a coherent and accurate prediction for the complex ICU patient.

There remain several avenues for further research. Ideally, TD learning should be evaluated on multiple architectures (e.g., Transformers, RNNs), with a sensitivity analysis to assess the impact of training dataset size. For instance, how does the potential for bias in TD learning evolve with smaller datasets, and at what sizes might it become significantly problematic? We aim to explore these questions in future work.

5. Conclusions

This report proposes a framework for applying temporal difference learning to outcome prediction tasks when using irregular time series health data. We subsequently implement this framework for the task of mortality estimation in the ICU, chosen for its partially observable data and long, complex trajectories. When compared to standard supervised learning methods, models trained with temporal difference learning can be shown to predict distant health outcomes with a higher level of accuracy,

and show greater resilience when evaluated on external unseen datasets. This work has important implications both for the implementation of temporal difference learning within healthcare reinforcement learning, and also for the wider field of health risk estimation.

Acknowledgments

We gratefully acknowledge the funding and facilities provided by the UKRI Centre for Doctoral Training in AI-enabled Healthcare (grant EP/S021612/1) and University College London respectively, which enabled this research. The views expressed in the text are those of the authors and do not necessarily reflect those of the above bodies.

Orcid

Thomas Frost https://orcid.org/0009-0002-5990-5800
Ken Li https://orcid.org/0000-0003-3073-3128
Steve Harris https://orcid.org/0000-0002-4982-1374

References

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

Maurizia Capuzzo, Carlo Alberto Volta, Tania Tassinati, Rui Paulo Moreno, Andreas Valentin, Bertrand Guidet, Gaetano Iapichino, Claude Martin, Thomas Perneger, Christophe Combescure, et al. Hospital mortality of adults admitted to intensive care units in hospitals with and without intermediate care units: a multicentre european cohort study. *Critical Care*, 18:1–15, 2014.

William Checkley, Greg S Martin, Samuel M Brown, Steven Y Chang, Ousama Dabbagh, Richard D Fremont, Timothy D Girard, Todd W Rice, Michael D Howell, Steven B Johnson, et al. Structure, process, and annual icu mortality across 69 centers: United states critical illness and injury trials group critical illness outcomes study. *Critical care medicine*, 42(2):344–356, 2014.

Farzad Esmaeili Tarki, Siamak Afaghi, Fatemeh Sadat Rahimi, Arda Kiani, Mohammad Varahram, and Atefeh Abedini. Serial sofa-score trends in icu-admitted covid-19 patients as predictor of 28-day mortality: A prospective cohort study. *Health Science Reports*, 6(5):e1116, 2023.

JE García-Gallo, NJ Fonseca-Ruiz, LA Celi, and JF Duitama-Muñoz. A machine learning-based model for 1-year mortality prediction in patients admitted to an intensive care unit with a diagnosis of sepsis. *Medicina intensiva*, 44(3):160–170, 2020.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Shinya Iwase, Taka-aki Nakada, Tadanaga Shimada, Takehiko Oami, Takashi Shimazui, Nozomi Takahashi, Jun Yamabe, Yasuo Yamao, and Eiryo Kawakami. Prediction algorithm for icu mortality and length of stay using machine learning. *Scientific reports*, 12(1):12912, 2022.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55, 2020.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Soo Yeon Kim, Saehoon Kim, Joongbum Cho, Young Suh Kim, In Suk Sol, Youngchul Sung, Inhyeok Cho, Minseop Park, Haerin Jang, Yoon Hee Kim, et al. A deep learning model for real-time mortality prediction in critically ill children. *Critical care*, 23:1–10, 2019.

Yeo Jin Kim, Markel Sanz Ausin, and Min Chi. Multi-temporal abstraction with time-aware deep q-learning for septic shock prevention. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1657–1663. IEEE, 2021.

Mihye Ko, Miyoung Shim, Sang-Min Lee, Yujin Kim, and Soyoung Yoon. Performance of apache iv in medical intensive care unit patients: comparisons with apache ii, saps 3, and mpm0 iii. *Acute and critical care*, 33(4):216–221, 2018.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Mingxing Lei, Zhencan Han, Shengjie Wang, Tao Han, Shenyun Fang, Feng Lin, and Tianlong Huang. A machine learning-based prediction model for in-hospital mortality among critically ill patients with hip fracture: An internal and external validated study. *Injury*, 54(2):636–644, 2023.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

TP Lillicrap. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Jiang Liu, Yihao Xie, Xin Shu, Yuwen Chen, Yizhu Sun, Kunhua Zhong, Hao Liang, Yujie Li, Chunyong Yang, Yan Han, et al. Value function assessment to different rl algorithms for heparin treatment policy of patients with sepsis in icu. *Artificial Intelligence in Medicine*, 147:102726, 2024.

Jayawant N Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.

Alexander Meyer, Dina Zverinski, Boris Pfahringer, Jörg Kempfert, Titus Kuehne, Simon H Sündermann, Christof Stamm, Thomas Hofmann, Volkmar Falk, and Carsten Eickhoff. Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet Respiratory Medicine*, 6(12):905–914, 2018.

Beatriz Nistal-Nuño. Developing machine learning models for prediction of mortality in the medical intensive care unit. *Computer Methods and Programs in Biomedicine*, 216:106663, 2022.

Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

Niklas Rodemund, Andreas Kokoefer, Bernhard Wernly, and Crispiana Cozowicz. Salzburg intensive care database (sicdb), a freely accessible intensive care database. *PhysioNet https://doi.org/10.13026/ezs8-6v88*, 2023a.

Niklas Rodemund, Bernhard Wernly, Christian Jung, Crispiana Cozowicz, and Andreas Koköfer. The salzburg intensive care database (sicdb): an openly available critical care dataset. *Intensive care medicine*, 49(6):700–702, 2023b.

Rahuldeb Sarkar, Christopher Martin, Heather Mattie, Judy Wawira Gichoya, David J Stone, and Leo Anthony Celi. Performance of intensive care unit severity scoring systems across different ethnicities in the usa: a retrospective observational study. *The Lancet Digital Health*, 3(4):e241–e249, 2021.

Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

Na Wang, Meiping Wang, Li Jiang, Bin Du, Bo Zhu, and Xiuming Xi. The predictive value of the oxford acute severity of illness score for clinical outcomes in patients with acute kidney injury. *Renal Failure*, 44(1):320–328, 2022a.

Xi Wang and Laurence Aitchison. How to set adamw's weight decay as you scale model and dataset size. *arXiv preprint arXiv:2405.13698*, 2024.

Yuqing Wang, Yun Zhao, and Linda Petzold. Predicting the need for blood transfusion in intensive care units with reinforcement learning. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10, 2022b.

Shu Zhou, Zongqing Lu, Yu Liu, Minjie Wang, Wuming Zhou, Xuanxuan Cui, Jin Zhang, Wenyan Xiao, Tianfeng Hua, Huaqing Zhu, et al. Interpretable machine learning model for early prediction of 28-day mortality in icu patients with sepsis-induced coagulopathy: development and validation. *European Journal of Medical Research*, 29(1):14, 2024.

Appendix A. Full Results

		Models										
Mortality	SOFA	*1d*	*1d*$^+$	*3d*	*3d*$^+$	*7d*	*7d*$^+$	*14d*	*14d*$^+$	*28d*	*28d*$^+$	*TD*
1-day	0.746	0.898	**0.901**	0.877	0.885	0.8521	0.860	0.831	0.842	0.827	0.833	0.874
3-days	0.718	0.838	**0.844**	0.831	0.840	0.814	0.821	0.799	0.811	0.800	0.807	0.841
7-days	0.700	0.806	0.816	0.813	0.821	0.802	0.804	0.787	0.800	0.798	0.803	**0.832**
14-days	0.697	0.789	0.800	0.802	0.809	0.796	0.795	0.780	0.795	0.799	0.803	**0.838**
28-days	0.694	0.780	0.788	0.793	0.801	0.785	0.785	0.770	0.786	0.793	0.796	**0.839**

Table 4: Mean AUROC results on the internal MIMIC-IV dataset (all models).
$^+$*indicates training with balanced cross-entropy.*

	Models										
Mortality	*1d*	*1d*$^+$	*3d*	*3d*$^+$	*7d*	*7d*$^+$	*14d*	*14d*$^+$	*28d*	*28d*$^+$	*TD*
1-day	0.860	**0.871**	0.839	0.838	0.810	0.824	0.789	0.801	0.776	0.783	0.865
3-days	0.782	0.799	0.776	0.772	0.755	0.765	0.732	0.748	0.728	0.734	**0.823**
7-days	0.720	0.741	0.726	0.720	0.716	0.720	0.694	0.712	0.704	0.708	**0.797**
14-days	0.687	0.712	0.701	0.695	0.695	0.696	0.675	0.695	0.693	0.697	**0.792**
28-days	0.664	0.692	0.682	0.676	0.680	0.677	0.658	0.681	0.680	0.688	**0.788**

Table 5: Mean AUROC results on the external SICdb dataset (all models).
$^+$*indicates training with balanced cross-entropy.*

Appendix B. Features

Category	Feature Name		
Antiarrhythmics	Amiodarone		
Antibiotics	Aciclovir	Ambisome	Amikacin
	Ampicillin	Ampicillin-Sulbactam	
	Azithromycin	Aztreonam	Caspofungin
	Cefazolin	Cefepime	Ceftazidime
	Ceftriaxone	Ciprofloxacin	Chloramphenicol
	Clindamycin	Colistin	Co-trimoxazole
	Daptomycin	Doxycycline	Ertapenem
	Erythromycin	Gentamicin	Levofloxacin
	Linezolid	Meropenem	Micafungin
	Metronidazole	Nafcillin	Oxacillin
	Piperacillin	Piperacillin-Tazobactam	
	Rifampin	Tigecycline	Tobramycin
	Vancomycin	Voriconazole	
Anticoagulants	Unfractionated Heparin		
Anticonvulsants	Levetiracetam	Phenytoin	
Antihypertensives	Nitroglycerin	Nitroprusside	
Beta-blockers	Labetalol		
Blood products	Fresh Frozen Plasma		Red Blood Cells

continued from previous page			
Category	**Feature Name**		
Blood products	Human Albumin Solution		Platelets
	IV Immune Globulin		
Diuretics	Bumetanide	Furosemide	
Glucose control	Regular Insulin		
Hyperosmotics	Hypertonic Saline 3%		Mannitol
Miscellaneous	Aminophylline	Sodium bicarbonate 8.4%	
Opioids	Fentanyl	Morphine sulphate	
Paralytics	Cisatracurium	Rocuronium	Vecuronium
Sedatives	Dexmedetomidine	Ketamine	Lorazepam
	Midazolam	Propofol	
Thrombolytics	Alteplase		
Vasopressors\Inotropes	Adrenaline	Dobutamine	Dopamine
	Milrinone	Noradrenaline	Vasopressin
Laboratory	Alanine transaminase		Albumin
	Alkaline phosphatase		Amylase
	Aspartate transferase		Anion gap
	Base excess	Bicarbonate	Blood O_2 pressure
	Blood CO_2 pressure		Blood pH
	Blood O_2 saturation		Calcium (ionised)
	Calcium (total)	C-reactive protein	Chloride
	Creatinine	Glucose (serum)	Glucose (bedside)
	Haematocrit	Haemoglobin	Lactate
	Lipase	Lactate dehydrogenase	
	Platelet count	Potassium	Prothrombin time
	Sodium	Total bilirubin	Troponin-T
	Urea	White blood cell count	
Demographics	Age	Gender	Weight

Table 6: Input Features

Appendix C. Additional Data Processing

In addition to the pre-processing steps described in the methods section, several additional steps were conducted at the feature level. Care was taken to ensure uniform units for all measurements and drug doses across both datasets. IV drugs were separated into bolus events (delivered over $\leq$1 minute) and absolute set rates. For example, a propofol infusion is started at 150mg/hr, with a 20mg bolus at 15 minutes, a rate increase to 200mg/hr after 30 minutes, and then stopped after 2 hours. If we treat the infusion ending as a state marker, this might be recorded as a 150mg/hr rate event at t=120, a 20mg bolus event at t=105, a 200mg/hr rate event at t=90, and a 0mg/hr rate event at t=0. Where two or more continuous infusions were given of the same drug, the rates were combined into a single value for any periods of overlap (i.e., the net rate of drug being delivered). Antibiotic scalar doses were excluded and the administration of an antibiotic was instead treated as a binary "bolus" event. When generating our model input data, the model can handle cases where there are fewer than 400 available measurements - however, when there were more than 400 measurements available in the eligible window, we prioritised 1) age/gender/weight, followed by 2) all current drug infusion rates, followed by 3) measurements ordered first by novelty (how many times has this feature already occurred), and then by decreasing recency. This latter step ensures that we prioritise including a broad range of features, and then newer information over older information. Outliers were removed according to the 0.005/0.995 quantiles of the training data for drug doses, and 0.001/0.999

quantiles of the training data for laboratory test results (per feature). In the rare cases where patient weight was never recorded, we imputed weight as the mean value for each gender according to the training data (74kg for females, 86kg for males). v and Δv were standardised to zero mean / unit variance for each feature individually (after outlier removal). t and Δt were standardised to zero mean / unit variance across all features.

Appendix D. Model Training Hyperparameters

The AdamW optimiser was used for each backpropagation update. The learning rate and weight decay were set as per Equation (5) and Equation (6) (Wang and Aitchison, 2024). Each model was trained for a maximum of 10 epochs, with the best iteration selected based on per-epoch evaluation on the validation data. The main and target networks were initialised with identical parameters at the start of training – the target network then receives a soft update of the main network's parameters after every update as per Equation (7), with $\alpha = 0.99$.

$$Learning\ rate = \frac{1}{n_{trainable\ parameters}} \tag{5}$$

$$Weight\ decay = \frac{1}{learning\ rate \cdot n_{batches\ per\ epoch}} \tag{6}$$

$$\theta_{target} \leftarrow \alpha * \theta_{target} + (1 - \alpha) * \theta_{main} \tag{7}$$