WRSE - a non-parametric weighted-resolution ensemble for predicting individual survival distributions in the ICU

Jonathan Heitz JONATHAN.HEITZ@INF.ETHZ.CH
Department of Computer Science, ETH Zürich, Zürich, Switzerland

Joanna Ficek JOANNA.FICEK@INF.ETHZ.CH
Department of Computer Science, ETH Zürich, Zürich, Switzerland
Life Science Zurich Graduate School, Zürich, Switzerland
Swiss Institute of Bioinformatics, Lausanne, Switzerland

Martin Faltys MARTIN.FALTYS@INF.ETHZ.CH
Department of Intensive Care Medicine, University Hospital, University of Bern, Bern, Switzerland
Department of Computer Science, ETH Zürich, Zürich, Switzerland

Tobias M. Merz TOBIASM@ADHB.GOVT.NZ
Department of Intensive Care Medicine, University Hospital, University of Bern, Bern, Switzerland
Cardiovascular Intensive Care Unit, Auckland City Hospital, Auckland, New Zealand

Gunnar Rätsch RAETSCH@INF.ETHZ.CH
Department of Computer Science, ETH Zürich, Zürich, Switzerland
Max Planck Institute for Intelligent Systems, Empirical Inference Department, Tübingen, Germany
Swiss Institute of Bioinformatics, Lausanne, Switzerland
University Hospital Zürich, Zürich, Switzerland

Matthias Hüser[*] MHUESER@INF.ETHZ.CH
Department of Computer Science, ETH Zürich, Zürich, Switzerland
Swiss Institute of Bioinformatics, Lausanne, Switzerland

Abstract

Dynamic assessment of mortality risk in the intensive care unit (ICU) can be used to stratify patients, inform about treatment effectiveness or serve as part of early-warning systems. Static risk scores, such as APACHE or SAPS, have been supplemented with data-driven approaches that track dynamic mortality risk over time. Recent works have focused on enhancing the information delivered to clinicians even further by producing full survival distributions instead of point predictions or fixed horizon risks. In this work, we propose a non-parametric ensemble model, *Weighted Resolution Survival Ensemble* (*WRSE*), tailored to estimate such dynamic individual survival distributions. Inspired by the simplicity and robustness of ensemble methods, the proposed approach combines a set of binary classifiers spaced according to a decay function reflecting the relevance of short-term predictions. Models and baselines are evaluated under weighted calibration and discrimination metrics for individual survival distributions, which closely reflect the utility of a model in ICU practice. We show competitive results with state-of-the-art probabilistic models, while greatly reducing training time by factors of 2-9x.

Keywords: intensive care unit, survival analysis, individual survival distribution

[*] To whom correspondence should be addressed

1. Introduction

Mortality prediction in the ICU has historically used scores, such as APACHE II (Knaus et al., 1985), which group patients into risk categories using data from the beginning of their stay (Keegan et al., 2011). For effective decision making, a more expressive risk estimate is desirable, such as predicting the full distribution over the remaining time-to-death (Avati et al., 2018; Haider et al., 2020). In this manner, individual patients with high or increasing mortality risk can be identified to direct the physician's attention. Decreasing mortality risk, on the other hand, can provide reassurance on treatment effectiveness.

To address this need, we propose a non-parametric approach that combines predictions of a set of classifiers using a weight function controlling the temporal spacing and hence, resolution in future time horizons. We call our approach *Weighted Resolution Survival Ensemble* (*WRSE*). By choosing decaying weight functions, our ensemble gives more importance to short-term predictions most relevant in various ICU settings. An example of *WRSE*'s output illustrating its use in ICU practice is shown in Fig. 1. We evaluate our model and baselines under weighted evaluation metrics that capture temporal calibration and discrimination.

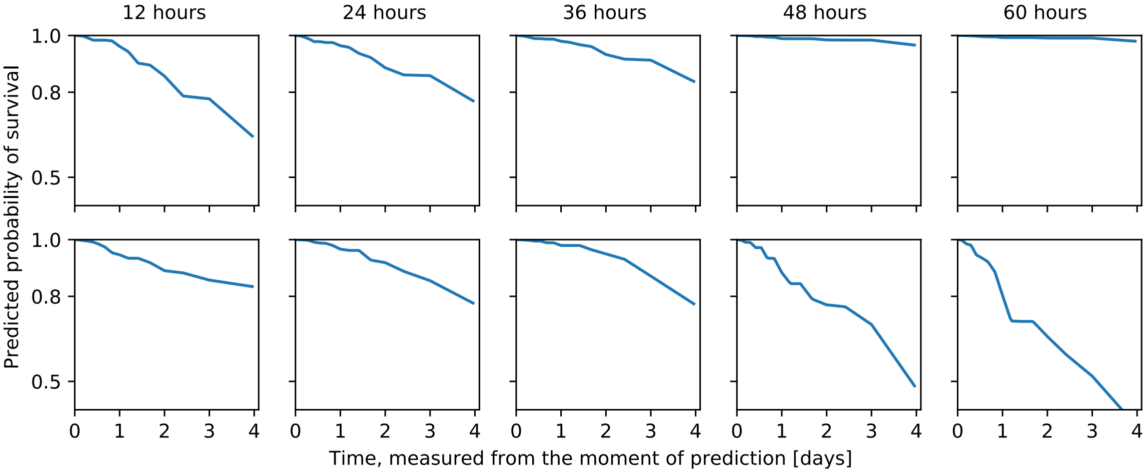

Figure 1: Each row displays the predicted dynamic survival curves for an example patient, estimated $\{12, 24, 36, 48, 60\}$ hours after admission to the ICU. According to *WRSE*'s estimations, the first patient's prognosis improves over time, while the second patient's curve indicates a worsening of prognosis.

2. Related work

The most commonly used approach for predicting survival curves, the Kaplan-Meier estimator (Kaplan and Meier, 1958), provides class-specific information on a population level, but cannot be used for producing individual survival distributions over time. To address this issue, several approaches, both parametric and non-parametric, have been proposed (Ishwaran et al., 2008; Yu et al., 2011; Avati et al., 2018). Furthermore, recent approaches

(Lee et al., 2018; Kvamme and Ørnulf Borgan, 2019) have leveraged neural networks to account for non-linear relationships and produce a full survival distribution per patient.

Several works have employed the strategy of combining different survival models to benefit from their strengths and obtain superior results to each model considered separately. Pirracchio et al. (2015) build a *Super Learner*, an ensemble of binary classifiers, to predict in-hospital death. This method, however, does not allow estimation of survival curves. Other approaches are based on *stacking*, i.e. combining either predictor matrices or predictions from survival models. In the first case, the predictor matrices of patients in the risk set at a given failure time point are concatenated together with the risk set indicator and a binary classifier is applied to the stacked matrix, yielding the conditional probability of experiencing the event at each considered time point (Zhong and Tibshirani, 2019). When combined with regression models, the approach can be used to estimate survival curves. In the second case, stacking survival models corresponds to providing a weighted combination of survival function estimates, with time-independent (Wey et al., 2015) or time-dependent weighting (Lee et al., 2019, *temporal quilting*). Such approaches are prone to overfitting because additional meta-parameters, more of them for time-dependent weighting, are introduced. Moreover, they suffer from the practical drawback of the training time being determined by the slowest base model.

On the topic of evaluation metrics for survival prediction, Cook (2006) argued that the key properties of a survival model are discrimination and calibration. Discrimination reflects the correct ordering of patients by the estimated probability of death. To take into account the time component, Antolini et al. (2005) introduced a *time-dependent discrimination index*, an adaptation of Harrell's C index (Harrell et al., 1982) that operates directly on predicted survival functions, instead of relying on point estimates. Calibration captures how well a model's predictions reflect the true frequency of the events. Haider et al. (2020) and Avati et al. (2018) generalized the notion of calibration to the full survival distribution. For both calibration and discrimination, recently proposed metrics ignore the relative importance of different future time horizons in an ICU. To bridge this gap, we propose weighted metrics that support selection of the best model for this particular application.

3. Dynamic individual survival distributions

We consider a temporal dataset $\{\{(\mathbf{x}_t^i, c^i, y_t^i)\}_{t=1}^{k_i}\}_{i=1}^N$ of N patients with (partially known) times-to-death T_t^i, and k_i the number of observed (hourly) time-points of patient i. Let $\mathbf{x}_t^i \in \mathbb{R}^d$ denote the feature vector, incorporating all information for a patient i at time t. Let c^i be the censoring indicator, with $c = 0$ denoting death at ICU. y_t^i is the observed time-to-death (if $c = 0$, then $T_t^i = y_t^i$) or time to discharge (if $c = 1$, then $T_t^i > y_t^i$). The latter may take place both in case of improvement (discharge to another hospital unit) and worsening (transfer to palliative care) of the patient's status and hence, the discharge leads to censoring. The task is to predict the distribution of the time-to-death T_t^i for all time-steps t during the ICU stay. We use $\hat{F}_t^i(\tau)$ to denote the predicted cumulative distribution function (CDF) for all future times $\tau > 0$, viewed from time point t. $\hat{S}_t^i(\tau) = 1 - \hat{F}_t^i(\tau)$ describes the predicted survival function.

4. Weighted evaluation metrics suitable for the ICU

To evaluate model performance, we use calibration and discrimination, which can be calculated for every future time point τ. To adapt these to better reflect clinical usefulness in an ICU, we weight short-term predictions more strongly, using an exponentially decaying weighting function $w(\tau) = \gamma^\tau$ with rate $\gamma \in (0, 1)$ and τ measured in days. $w(\tau) = \gamma^\tau$ is plotted for $\gamma = 0.3$, $\gamma = 0.5$ and $\gamma = 0.8$ in Fig. 2. The three settings are examples of strong weighting of the short-term future (next 2 days), as well as two more moderate decays. Depending on the desired application and its time horizon, other choices of γ may be appropriate. We evaluate calibration of the estimated $Pr[T < \tau]$ for every time τ, plotting binned mean predicted values against the fraction of positives. This produces a piece-wise linear function $c_\tau(q)$ with $c_\tau(0) = 0$ and $c_\tau(1) = 1$ and the diagonal $c^*(q) = q$ representing perfect calibration. We report the absolute area around the diagonal as a metric of calibration for every τ: $\int_0^1 |c_\tau(q) - c^*(q)|\, dq$. Patients who do not die are not considered for τ after their time of discharge. The *weighted absolute area around the diagonal* Cal^w is then given as a weighted mean:

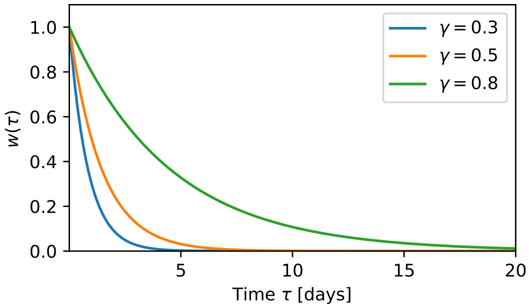

Figure 2: The exponentially decaying weighting function $w(\tau) = \gamma^\tau$.

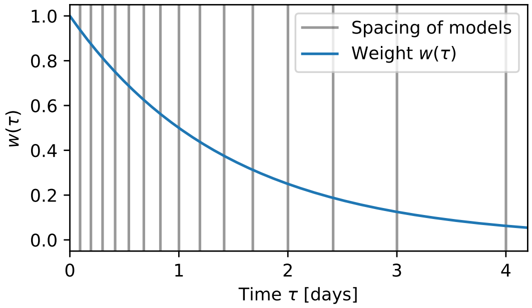

Figure 3: Spacing of 15 base models in *WRSE* ensemble weighted according to $\gamma = 0.5$ (cf. Section 5.1)

$$Cal^w = \frac{\sum_{\tau \in T} w(\tau) \cdot \int_0^1 |c_\tau(q) - c^*(q)|\, dq}{\sum_{\tau \in T} w(\tau)} .$$

To evaluate discrimination, we adapt the time-dependent discrimination index C^{td} (Antolini et al., 2005). Let $\mathcal{T}$ be the set of distinct times of death ($\mathcal{T} = \{y^k | c^k = 0\}$). For each $\tau \in \mathcal{T}$, we consider the set of pairs of patients $\mathcal{P}_\tau = \{(i, j) | y^i = \tau \wedge c^i = 0 \wedge y^j \geq \tau\}$. Here, patient i has died at time τ and patient j is still at risk, i.e. has neither died nor been discharged until time τ. A pair is concordant if $\hat{S}^i(y^i) \leq \hat{S}^j(y^i)$ and C^{td} averages the fraction of concordant pairs over time. The *weighted time-dependent discrimination index* $C^{td,w}$ is then:

$$C^{td,w} = \frac{\sum_{\tau \in \mathcal{T}} |\{(i, j) \in \mathcal{P}_\tau | \hat{S}^i(\tau) < \hat{S}^j(\tau)\}| \cdot w(\tau)}{\sum_{\tau \in \mathcal{T}} |\mathcal{P}_\tau| \cdot w(\tau)} .$$

5. WRSE and baseline models

5.1. WRSE (Weighted Resolution Survival Ensemble)

The proposed non-parametric ensemble estimator (*WRSE*) consists of binary classification base models $m_1, \ldots, m_K$, where $m_k(\mathbf{x})$ for $k \in \{1, \ldots, K\}$ predicts the probability of dying within the next h_k hours ($Pr[T < h_k \mid \mathbf{x}]$), given patient information $\mathbf{x}$ and an increasing sequence $h = (h_1, \ldots, h_K)$. We use $K = 15$ and define $h_k = w^{-1}(1 - \frac{k}{K+1})$, where $w^{-1}(\cdot) = \log(\cdot)/\log(\gamma)$ is the inverse of $w(\cdot)$ defined in Section 4, thereby putting more emphasis on the short-term future, as depicted in Fig. 3. We combine the base models' predictions by interpreting them as non-parametric estimates of the cumulative distribution function (CDF) for $0 \leq T \leq h_K$. We do not predict the shape of the CDF for $T > h_K$, as such long-term horizons are rarely present in the ICU. Since the base models are independent, the sequence $m_1(\mathbf{x}), \ldots, m_K(\mathbf{x})$ is not guaranteed to be monotonically increasing for a given patient, a necessary condition for a CDF. We solve this using the isotonic regression framework (Barlow et al., 1972), finding an optimal fit vector $m_1^*, \ldots, m_K^*$ subject to monotonicity constraints. More precisely, isotonic regression solves the following constrained optimization problem:

$$\min \sum_{k=1}^{K} (m_k^* - m_k(\mathbf{x}))^2$$
$$\text{subject to } m_1^* \leq m_2^* \leq \cdots \leq m_K^* \ .$$

The ensemble framework is general, and any binary classification model could be used. We employ *LightGBM* because of its demonstrated high performance for ICU data (Hyland et al., 2020) and interpretability using the TreeSHAP algorithm Lundberg et al. (2018). Each *LightGBM* base model has at most 64 leaves in each tree, at most 1000 trees, and a learning rate of 0.01.

5.2. Parametric and non-parametric baselines

As parametric baselines, we evaluate two models within the Survival-CRPS framework (Avati et al., 2018): *Log-normal*, for its flexibility (Royston, 2001; Yang et al., 2017) and *Exponential*, for its good fit to our ICU data-set seen in a preliminary analysis. A detailed description of the framework and implementation details can be found in Appendix Section A.2.1. As non-parametric baselines, which avoid strong assumptions about the underlying distribution, we use (1) *DeepHit* (Lee et al., 2018), a model discretizing time into intervals and jointly predicting the probability of dying in each interval, using a likelihood loss combined with a ranking loss function, (2) *Nnet-survival* (also called *Logistic-Hazard* (Kvamme and Ørnulf Borgan, 2019)) by Gensheimer and Narasimhan (2019), which predicts the conditional probability of dying within each interval, using a custom likelihood loss function, and (3) multi-task logistic regression (*MTLR*) (Yu et al., 2011), which is based on a likelihood loss function optimizing a set of dependent logistic regressors for future time points. Implementation details and hyperparameter settings for the non-parametric baselines can be found in Appendix Section A.2.2.

6. Experiments

6.1. Data set and evaluation setup

We use the HiRID data set (Faltys et al., 2020), containing time series of more than 50,000 admissions to a tertiary-care ICU. The data includes organ function parameters, lab results, and treatment parameters. The 30 most important variables, given by the highest mean absolute SHAP values (Lundberg and Lee, 2017) on the validation set, were used for model construction and are listed in Appendix Table 5. Missing values were imputed using forward filling and filling with a clinically normal value if no measurement was present prior to a time-point. We drew 5 replicates of splits, each consisting of a training, a validation and a test set (more details are provided in the Appendix Section A.1). Patients in the test set have a later admission time than patients in the training set, simulating model deployment on future data. We train our models on the training set of each split. The validation set is used for early stopping, selection of optimal hyperparameters using grid search, and the analysis of variable importance. We evaluate models on the test set, considering patients once every hour as independent test instances to estimate the future survival distribution. We refer to Hyland et al. (2020) for details on preprocessing and the temporal splits.

6.2. Comparison with baselines

WRSE (with spacing γ=0.5) and the baseline models were compared under the two weighted metrics described in Section 4. All baselines were recalibrated using isotonic regression. *WRSE* did not require this step, as raw predictions already show sufficient calibration. The results in Table 1 demonstrate that *WRSE* outperforms parametric baselines and is on par with state-of-the-art non-parametric approaches, with respect to discrimination ($C^{td,w}$). The trend persists across different weighting functions. Furthermore, it is better calibrated than all baselines for γ=0.3, which up-weights short-term predictions, while being outperformed by *DeepHit* for γ=0.8. We complement these results by plotting the unweighted raw metrics against the time horizon τ in Fig. 4 (discrimination) and in Fig. 5 (calibration). *WRSE* consists of a set of independent classifiers, which can be trained in parallel, with isotonic regression applied subsequently. We observe reduced training times by a factor of 2-4x (parametric baselines), and 5-9x (non-parametric baselines), as displayed in Table 2.

Table 1: Results of our model compared with the baselines, listing the metrics described in Section 4 with three evaluation weighting functions ($\gamma \in \{0.3, 0.5, 0.8\}$). We report mean and standard deviation across the 5 splits.

Model	$C^{td,w}$, $\gamma = 0.3$	$C^{td,w}$, $\gamma = 0.5$	$C^{td,w}$, $\gamma = 0.8$	Cal^{w}, $\gamma = 0.3$	Cal^{w}, $\gamma = 0.5$	Cal^{w}, $\gamma = 0.8$
Log-normal	0.78 ± 0.02	0.77 ± 0.02	0.75 ± 0.01	0.17 ± 0.05	0.15 ± 0.05	0.12 ± 0.04
Exponential	0.87 ± 0.02	0.85 ± 0.02	0.83 ± 0.02	0.13 ± 0.05	0.14 ± 0.04	0.12 ± 0.03
DeepHit	**0.91** ± 0.01	**0.89** ± 0.01	**0.87** ± 0.01	0.11 ± 0.02	**0.09** ± 0.01	**0.06** ± 0.01
Logistic Hazard	0.90 ± 0.01	**0.89** ± 0.01	0.86 ± 0.01	0.11 ± 0.02	0.10 ± 0.02	0.09 ± 0.02
MTLR	**0.91** ± 0.01	**0.89** ± 0.01	**0.87** ± 0.01	0.16 ± 0.03	0.12 ± 0.02	0.08 ± 0.01
WRSE (ours)	**0.92** ± 0.01	**0.90** ± 0.01	**0.88** ± 0.01	**0.07** ± 0.01	**0.08** ± 0.01	0.09 ± 0.01

Table 2: Average training time of *WRSE* (parallelized ensemble trained on a multi-core CPU) and the baselines (trained on one GPU), across the 5 splits.

Model	Log-normal	Exponential	DeepHit	Logistic Hazard	MTLR	WRSE (ours)
Training time [min]	283	127	420	349	588	**63**

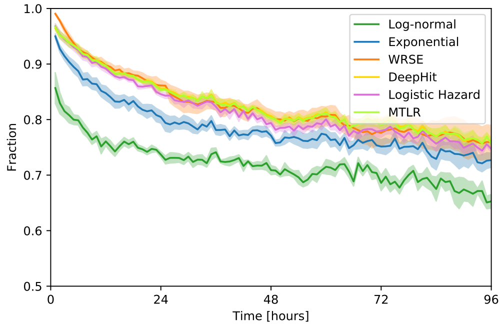

Figure 4: Fraction of concordant pairs per horizon τ (cf. Section 4). Higher values represent better discrimination. Shaded error bands denote the standard error across the 5 splits.

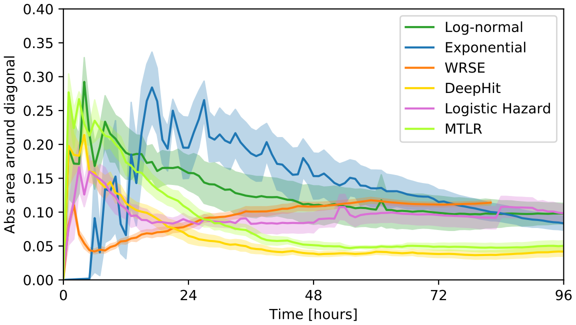

Figure 5: Area around the diagonal of the calibration plot in each horizon τ. Lower values represent better calibration. Shaded error bands denote the standard error across the 5 splits.

6.3. Analyzing different WRSE configurations

To understand the effect of different weighting schemes on *WRSE* performance, we analyze several versions, varying the number of base models (5, 7, and 10) and their spacing in time (evenly spaced vs. weighted spacing). We observe that the weighted versions exhibit superior calibration for short-term horizons, and for long-term horizons if the number of base models is small (Table 3). The discrimination performance of all variants is similar. More detailed results are displayed in Appendix A.3.2. We further analyse alternative choices of base models (a multi-layer perceptron and logistic regression) and observe superior performance of LightGBM across various *WRSE* configurations, as discussed in Appendix Section A.3.3.

Table 3: Results contrasting *WRSE* with temporal weighting of base models according to $\gamma = 0.5$ (cf. Section 5.1) and evenly spaced base models covering 10 days. We report the mean and standard deviation across the 5 splits.

Model	$C^{td,w}$, $\gamma = 0.3$	$C^{td,w}$, $\gamma = 0.5$	$C^{td,w}$, $\gamma = 0.8$	Cal^{w}, $\gamma = 0.3$	Cal^{w}, $\gamma = 0.5$	Cal^{w}, $\gamma = 0.8$
Even spacing 5 models	0.91 ± 0.01	0.90 ± 0.01	0.87 ± 0.01	0.11 ± 0.01	0.11 ± 0.01	0.10 ± 0.01
Weighted $\gamma = 0.5$, 5 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	**0.07** ± 0.01	**0.08** ± 0.01	**0.08** ± 0.01
Even spacing 7 models	0.91 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.10 ± 0.01	0.10 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.5$, 7 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	**0.08** ± 0.02	**0.08** ± 0.01	0.09 ± 0.01
Even spacing 10 models	0.91 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.10 ± 0.01	0.10 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.5$, 10 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	**0.07** ± 0.01	**0.08** ± 0.01	0.09 ± 0.01

6.4. Performance of WRSE in different diagnostic groups

We further analyze the calibration and discrimination of *WRSE* in sub-cohorts corresponding to diagnostic groups. We only include sub-cohorts with at least 100 patients in the test set, whose characteristics are shown in Appendix Table 6. The results, displayed in Table 4, indicate that the performance is good across all the groups. Furthermore, the discrimination is highest for neurological and trauma patients, whereas calibration is best for neurological and cardiovascular patients.

Table 4: Performance of *WRSE* for different patient sub-cohorts in the test set of HiRID. We report the mean and standard deviation across the 5 splits. Only diagnostic groups with at least 100 patients in the test set were included. Diagnostic group size is indicated in parentheses.

Diagnostic group	$C^{td,w}$, $\gamma = 0.3$	$C^{td,w}$, $\gamma = 0.5$	$C^{td,w}$, $\gamma = 0.8$	Cal^{w}, $\gamma = 0.3$	Cal^{w}, $\gamma = 0.5$	Cal^{w}, $\gamma = 0.8$
Neurologic (n=1006)	0.94 ± 0.01	0.93 ± 0.01	0.90 ± 0.02	0.07 ± 0.01	0.08 ± 0.02	0.09 ± 0.02
Cardiovascular (n=949)	0.91 ± 0.02	0.89 ± 0.02	0.87 ± 0.02	0.09 ± 0.01	0.09 ± 0.01	0.10 ± 0.02
Gastrointestinal (n=516)	0.90 ± 0.03	0.89 ± 0.03	0.87 ± 0.04	0.09 ± 0.01	0.10 ± 0.01	0.11 ± 0.02
Respiratory (n=354)	0.89 ± 0.01	0.88 ± 0.02	0.85 ± 0.02	0.10 ± 0.02	0.11 ± 0.03	0.11 ± 0.04
Trauma (n=241)	0.93 ± 0.03	0.92 ± 0.03	0.89 ± 0.05	0.11 ± 0.02	0.12 ± 0.02	0.13 ± 0.03

7. Discussion

We presented *Weighted Resolution Survival Ensemble* (*WRSE*), a non-parametric method that estimates dynamic individual survival predictions in the ICU. In its default setting (γ=0.5), it allocates more classifiers for short-term predictions, which are more relevant in several scenarios in an ICU. Comparisons against various parametric and non-parametric baselines show similar or superior performance. Our framework is adaptive to the user's choices via its weighting function, and the choice of base models. Once the temporal spacing of *WRSE* is set, the base models are trivially parallelizable, resulting in a training time decrease of 2-9 times compared to the baselines. We also proposed modifications to time-dependent calibration and discrimination metrics up-weighting short-term predictions. We believe that these metrics capture a model's usefulness in a dynamic ICU setting more closely. Future work will focus on evaluation in other cohorts, approaches for deciding the optimal temporal spacing, given a fixed budget of base models, as well as the introduction of other base models for improved uncertainty estimation.

Acknowledgments

This project was supported by the grant #2017-110 of the Strategic Focus Area ”Personalized Health and Related Technologies (PHRT)” of the ETH Domain. JF/MH are partially supported by ETH core funding (to GR). MF/MH are supported by the Grant No. 205321 176005 of the Swiss National Science Foundation (to GR). The authors would like to thank Ximena Bonilla for providing useful feedback on the manuscript.

References

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

Anand Avati, Tony Duan, Kenneth Jung, Nigam H Shah, and Andrew Ng. Countdown regression: sharp and calibrated survival predictions. *arXiv preprint arXiv:1806.08324*, 2018.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. J. Wiley, 1972. ISBN 9780471049708.

D. A. Cook. Methods to assess performance of models estimating risk of death in intensive care patients: A review. *Anaesthesia and Intensive Care*, 34(2):164–175, 2006. doi: 10.1177/0310057X0603400205.

Martin Faltys, Marc Zimmermann, Xinrui Lyu, Matthias Hüser, Stephanie Hyland, Gunnar Rätsch, and Tobias Merz. Hirid, a high time-resolution icu dataset (version 1.0), 2020.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. In *Advances in Neural Information Processing Systems*, pages 4650–4661, 2019.

Michael F Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85): 1–63, 2020.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373, 2020.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *Ann. Appl. Stat.*, 2(3):841–860, 09 2008. doi: 10.1214/08-AOAS169.

E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 2020/09/29/ 1958. doi: 10.2307/2281868.

Mark T. Keegan, Ognjen Gajic, and Bekele Afessa. Severity of illness scoring systems in the intensive care unit. *Critical Care Medicine*, 39(1), 2011.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL http://arxiv.org/abs/1412.6980. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks, 2019.

Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *AAAI*, pages 2314–2321, 2018.

Changhee Lee, William Zame, Ahmed Alaa, and Mihaela van der Schaar. Temporal quilting for survival analysis. volume 89 of *Proceedings of Machine Learning Research*, pages 596–605. PMLR, 16–18 Apr 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Romain Pirracchio, Maya L Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, and Mark J van der Laan. Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study. *The Lancet Respiratory Medicine*, 3(1):42 – 52, 2015. ISSN 2213-2600. doi: https://doi.org/10.1016/S2213-2600(14)70239-5.

P. Royston. The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Statistica Neerlandica*, 55(1):89–104, 2001. doi: 10.1111/1467-9574.00158.

Andrew Wey, John Connett, and Kyle Rudser. Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics*, 16(3):537–549, 02 2015. ISSN 1465-4644. doi: 10.1093/biostatistics/kxv001.

Yinchong Yang, Peter A. Fasching, and Volker Tresp. Modeling progression free survival in breast cancer with tensorized recurrent neural networks and accelerated failure time models. volume 68 of *Proceedings of Machine Learning Research*, pages 164–176, Boston, Massachusetts, 18–19 Aug 2017. PMLR.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1845–1853. Curran Associates, Inc., 2011.

Chenyang Zhong and Robert Tibshirani. Survival analysis as a classification problem. *arXiv preprint arXiv:1909.11171*, 2019.

Appendix A. Appendix

A.1. Data set details

The HiRID data set (Faltys et al., 2020) used in this work contains information about more than 50,000 ICU stays collected between 2008 and 2016. Table 5 displays the 30 (of a total of 197) variables used in the experiments. To train models and evaluate performance, 5 partly overlapping temporal splits have been drawn. Each split covers roughly six years, of which the first 5/6 are used for training, 1/12 for validation and 1/12 for testing. This setup simulates deployment to future data. We refer the reader to Hyland et al. (2020) for detailed information on the data collection, preprocessing, and the resulting splits. Table 6 shows the clinical characteristics of patients in the test set, per diagnostic group with at least 100 patients.

Table 5: The subset of variables from the HiRID data set used in our experiments, in parentheses are the meta-variable IDs of the parameters as defined in Hyland et al. (2020).

Organ function parameters
PAPm (vm9)
Cardiac output (vm13)
Heart rhythm state (vm19)
SpO_2 (vm20)
Respiratory rate (vm22)
Supplemental oxygen (vm23)
Urine output / time (vm24)
GCS Verbal / Motor / Eye (vm25-27)
RASS (vm28)
Fluid output / time (vm32)
FiO_2 (vm58)
Weight (vm131)
Age
Treatment parameters
Norepinephrine (pm39)
Ventilator mode / Peak pressure (vm60,62)
Ventilator RR setting (vm65)
Propofol (pm80)
Hourly CSF drainage (vm84)
Steroids (pm91)
Laboratory tests
Arterial lactate (vm136)
Creatine kinase (vm144)
Mg (vm154)
Urea (vm155)
Bilirubine, total (vm162)
aPTT (vm166)
Total white blood cell count (vm184)
Platelet count (vm185)

Table 6: Number of test set patients and test set instances per diagnostic group with at least 100 patients in the test set. We report the mean and standard deviation across the 5 splits.

Diagnostic group	# Patients	Mortality rate [%]	# Test instances	Positive prevalence [%]
Cardiovascular	949 ± 117	6.5 ± 0.9	35382 ± 2212	10.1 ± 2.0
Neurologic	1006 ± 105	5.9 ± 0.7	42816 ± 2204	7.3 ± 2.3
Respiratory	354 ± 39	7.0 ± 0.5	15475 ± 1442	9.4 ± 2.8
Gastrointestinal	516 ± 89	6.6 ± 0.7	19475 ± 2444	8.7 ± 1.7
Trauma	241 ± 37	7.1 ± 1.6	10798 ± 1632	7.0 ± 1.5

A.2. Implementation details

A.2.1. Parametric baselines

The Survival-CRPS framework (Avati et al., 2018) allows to incorporate information from censored observations into the model, with the loss function given by

$$\mathcal{S}_{CRPS}\left(\hat{F}_t,(y,c)\right)=\int_0^y \hat{F}_t(\tau)^2\,d\tau+(1-c)\int_y^\infty\left(1-\hat{F}_t(\tau)\right)^2\,d\tau\ .$$

For the log-normal distribution, a closed-form representation of $\mathcal{S}_{CRPS}$ does not exist and thus, we use a trapezoidal approximation suitable for backpropagation (Avati et al., 2018). We derive the Survival-CRPS loss function for the exponential distribution. The cumulative density function is given by $\hat{F}_t^{\lambda}(\tau)=1-\exp(-\lambda\cdot\tau)$ for a parameter λ. $\mathcal{S}_{CRPS}$ then has a closed-form representation

$$\begin{aligned}&\mathcal{S}_{CRPS}\left(\hat{F}_t^{\lambda},(y,c)\right)\\&=\int_0^y \hat{F}_t^{\lambda}(\tau)^2\,d\tau+(1-c)\int_y^\infty\left(1-\hat{F}_t^{\lambda}(\tau)\right)^2\,d\tau\\&=\int_0^y(1-e^{-\lambda\tau})^2\,d\tau+(1-c)\int_y^\infty e^{-2\lambda\tau}\,d\tau\\&=\frac{4e^{-\lambda y}-ce^{-2\lambda y}-3}{2\lambda}+y\ .\end{aligned}$$

We follow the implementation of Survival-CRPS provided by the authors (Avati et al., 2018). All models were trained using an Adam optimizer (Kingma and Ba, 2014) and early stopping, terminating training when the validation loss does not improve for 10 epochs. Two hidden layers of 50 neurons each, moderate weight regularization of 0.01, and a learning rate of 1e-4 was used. The output layer directly predicts one (exponential) or two (log-normal) distribution parameters. We analyze two feature extractors: an MLP with two hidden layers and a temporal convolutional network, shown to outperform RNNs in sequence modeling (Bai et al., 2018; Franceschi et al., 2019). Results contrasting the different feature extractor choices for the parametric baselines are discussed in Section A.3.1.

A.2.2. Non-parametric baselines

For *DeepHit*, *Logistic-Hazard* and *MTLR*, we use an implementation using the `pycox` library. Specific details about the implementations are given in Kvamme et al. (2019), and Kvamme and Ørnulf Borgan (2019), respectively. For *DeepHit* we use three layers with 240, 400, and 240 nodes, as well as parameters $\alpha = 0.4$ and $\sigma = 2$, and a learning rate of 3.3e-07. For *Logistic-Hazard* we use a neural network consisting of three fully-connected layers with 90, 150, and 90 nodes and a learning rate of 3.3e-07. *MTLR* uses a neural network consisting of three fully-connected layers with 90, 150, and 90 nodes and a learning rate of 3.3e-07. *DeepHit*, *Logistic-Hazard*, and *MTLR* are discrete-time models, requiring discretization of continuous times into intervals. We use 56 intervals, each covering roughly 12h. All models were trained using an Adam optimizer (Kingma and Ba, 2014) and early stopping, terminating training when the loss on the validation set does not improve for 10 epochs.

A.3. Additional results

A.3.1. Feature extractors for parametric baseline models

Results contrasting the different feature extractor choices for the parametric baselines are shown in Table 7. The TCN-based log-normal model shows better calibration and discrimination than its MLP-based counterpart. However, calibration is clearly worse. As a result, we consider the MLP-based model superior, as uncalibrated results are useless in practice. The exponential models are very similar; for consistency and simplicity, we define the MLP-based version as our model of choice to be included in the main results in Section 6.2.

Table 7: Performance of parametric models (log-normal and exponential) with two different feature extractors. The first is a multi-layer perceptron (MLP) on the current time-point. As a second feature extractor, we use a temporal convolution network (TCN) architecture on 24h of patient history. We report the mean and standard deviation across the 5 splits.

Model	$C^{td,w}$, $\gamma = 0.3$	$C^{td,w}$, $\gamma = 0.8$	Cal^{w}, $\gamma = 0.3$	Cal^{w}, $\gamma = 0.8$
Log-normal MLP	0.78 ± 0.02	0.75 ± 0.01	**0.16** ± 0.05	**0.12** ± 0.04
Log-normal TCN	**0.89** ± 0.03	**0.83** ± 0.02	0.21 ± 0.01	**0.14** ± 0.01
Exponential MLP	**0.87** ± 0.02	**0.83** ± 0.02	**0.12** ± 0.04	**0.12** ± 0.03
Exponential TCN	**0.88** ± 0.02	**0.84** ± 0.03	**0.16** ± 0.03	**0.11** ± 0.01

A.3.2. Effect of weighted temporal spacing

An analysis of the effect of weighted temporal spacing on the discrimination index by time horizon τ is shown in Fig. 7. It is apparent that temporal spacing has no effect on temporal discrimination, in contrast to calibration (Fig. 6), as shown in the main paper.

A.3.3. Alternative base models

Table 9 contrasts the performance of *WRSE* for different base models: (1) LightGBM (proposed model), (2) a multi-layer perceptron (MLP) and (3) logistic regression. Choosing LightGBM as base model leads to superior or comparable results across varying number of models and weighting functions used.

Table 8: Results contrasting *WRSE* with different temporal spacing of base models. Metrics described in Section 4 with three weighing functions ($\gamma \in \{0.3, 0.5, 0.8\}$) are shown. We report the mean and standard deviation across the 5 splits.

Model	$C^{td,w}$, $\gamma = 0.3$	$C^{td,w}$, $\gamma = 0.5$	$C^{td,w}$, $\gamma = 0.8$	Cal^w, $\gamma = 0.3$	Cal^w, $\gamma = 0.5$	Cal^w, $\gamma = 0.8$
Even spacing 5 models	0.91 ± 0.01	0.90 ± 0.01	0.87 ± 0.01	0.11 ± 0.01	0.11 ± 0.01	0.10 ± 0.01
Weighted $\gamma = 0.3$, 5 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.07 ± 0.01	0.07 ± 0.01	0.07 ± 0.01
Weighted $\gamma = 0.5$, 5 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.07 ± 0.01	0.08 ± 0.01	0.08 ± 0.01
Weighted $\gamma = 0.8$, 5 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.09 ± 0.01	0.09 ± 0.01	0.09 ± 0.01
Even spacing 7 models	0.91 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.10 ± 0.01	0.10 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.3$, 7 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.07 ± 0.01	0.07 ± 0.01	0.07 ± 0.01
Weighted $\gamma = 0.5$, 7 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.08 ± 0.02	0.08 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.8$, 7 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.09 ± 0.01	0.09 ± 0.01	0.09 ± 0.01
Even spacing 10 models	0.91 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.10 ± 0.01	0.10 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.3$, 10 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.07 ± 0.01	0.07 ± 0.01	0.08 ± 0.01
Weighted $\gamma = 0.5$, 10 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.07 ± 0.01	0.08 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.8$, 10 models	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.08 ± 0.01	0.08 ± 0.01	0.09 ± 0.01

Table 9: Results for *WRSE* with LightGBM (proposed model) and two alternative base models: A multi-layer perceptron (MLP) and logistic regression. We report the mean and standard deviation across the 5 splits.

Model	$C^{td,w}$, $\gamma = 0.3$	$C^{td,w}$, $\gamma = 0.5$	$C^{td,w}$, $\gamma = 0.8$	Cal^w, $\gamma = 0.3$	Cal^w, $\gamma = 0.5$	Cal^w, $\gamma = 0.8$
Weighted $\gamma = 0.3$, 5 models (LightGBM)	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.07 ± 0.01	0.07 ± 0.01	0.07 ± 0.01
Weighted $\gamma = 0.3$, 5 models (MLP)	0.89 ± 0.02	0.88 ± 0.02	0.85 ± 0.01	0.11 ± 0.01	0.11 ± 0.01	0.10 ± 0.01
Weighted $\gamma = 0.3$, 5 models (Logistic Regression)	0.88 ± 0.01	0.87 ± 0.01	0.85 ± 0.01	0.13 ± 0.03	0.13 ± 0.03	0.12 ± 0.03
Weighted $\gamma = 0.5$, 5 models (LightGBM)	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.07 ± 0.01	0.08 ± 0.01	0.08 ± 0.01
Weighted $\gamma = 0.5$, 5 models (MLP)	0.89 ± 0.01	0.88 ± 0.01	0.85 ± 0.01	0.11 ± 0.01	0.11 ± 0.01	0.10 ± 0.00
Weighted $\gamma = 0.5$, 5 models (Logistic Regression)	0.89 ± 0.01	0.88 ± 0.01	0.85 ± 0.01	0.11 ± 0.01	0.10 ± 0.01	0.10 ± 0.01
Weighted $\gamma = 0.8$, 5 models (LightGBM)	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.09 ± 0.01	0.09 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.8$, 5 models (MLP)	0.90 ± 0.01	0.89 ± 0.01	0.86 ± 0.01	0.11 ± 0.04	0.11 ± 0.03	0.11 ± 0.02
Weighted $\gamma = 0.8$, 5 models (Logistic Regression)	0.90 ± 0.01	0.88 ± 0.01	0.86 ± 0.01	0.13 ± 0.04	0.12 ± 0.03	0.12 ± 0.02
Weighted $\gamma = 0.3$, 7 models (LightGBM)	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.07 ± 0.01	0.07 ± 0.01	0.07 ± 0.01
Weighted $\gamma = 0.3$, 7 models (MLP)	0.88 ± 0.02	0.87 ± 0.01	0.85 ± 0.01	0.12 ± 0.02	0.12 ± 0.02	0.11 ± 0.02
Weighted $\gamma = 0.3$, 7 models (Logistic Regression)	0.88 ± 0.01	0.87 ± 0.01	0.85 ± 0.01	0.12 ± 0.02	0.12 ± 0.02	0.11 ± 0.02
Weighted $\gamma = 0.5$, 7 models (LightGBM)	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.08 ± 0.02	0.08 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.5$, 7 models (MLP)	0.89 ± 0.02	0.88 ± 0.02	0.86 ± 0.01	0.11 ± 0.02	0.10 ± 0.02	0.10 ± 0.02
Weighted $\gamma = 0.5$, 7 models (Logistic Regression)	0.88 ± 0.02	0.87 ± 0.02	0.85 ± 0.01	0.12 ± 0.04	0.11 ± 0.03	0.10 ± 0.03
Weighted $\gamma = 0.8$, 7 models (LightGBM)	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.09 ± 0.01	0.09 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.8$, 7 models (MLP)	0.90 ± 0.01	0.89 ± 0.01	0.86 ± 0.01	0.11 ± 0.03	0.11 ± 0.02	0.12 ± 0.03
Weighted $\gamma = 0.8$, 7 models (Logistic Regression)	0.90 ± 0.01	0.88 ± 0.01	0.86 ± 0.01	0.12 ± 0.03	0.11 ± 0.02	0.12 ± 0.02
Weighted $\gamma = 0.3$, 10 models (LightGBM)	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.07 ± 0.01	0.07 ± 0.01	0.08 ± 0.01
Weighted $\gamma = 0.3$, 10 models (MLP)	0.88 ± 0.01	0.87 ± 0.01	0.85 ± 0.01	0.10 ± 0.01	0.10 ± 0.01	0.10 ± 0.01
Weighted $\gamma = 0.3$, 10 models (Logistic Regression)	0.88 ± 0.01	0.87 ± 0.01	0.85 ± 0.01	0.11 ± 0.02	0.11 ± 0.02	0.10 ± 0.01
Weighted $\gamma = 0.5$, 10 models (LightGBM)	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.07 ± 0.01	0.08 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.5$, 10 models (MLP)	0.89 ± 0.01	0.88 ± 0.01	0.86 ± 0.01	0.11 ± 0.01	0.11 ± 0.01	0.10 ± 0.01
Weighted $\gamma = 0.5$, 10 models (Logistic Regression)	0.88 ± 0.02	0.87 ± 0.02	0.85 ± 0.01	0.10 ± 0.01	0.10 ± 0.02	0.10 ± 0.02
Weighted $\gamma = 0.8$, 10 models (LightGBM)	0.92 ± 0.01	0.90 ± 0.01	0.88 ± 0.01	0.08 ± 0.01	0.08 ± 0.01	0.09 ± 0.01
Weighted $\gamma = 0.8$, 10 models (MLP)	0.90 ± 0.01	0.88 ± 0.01	0.86 ± 0.01	0.11 ± 0.02	0.11 ± 0.02	0.12 ± 0.02
Weighted $\gamma = 0.8$, 10 models (Logistic Regression)	0.90 ± 0.01	0.88 ± 0.01	0.86 ± 0.01	0.11 ± 0.01	0.11 ± 0.01	0.12 ± 0.02

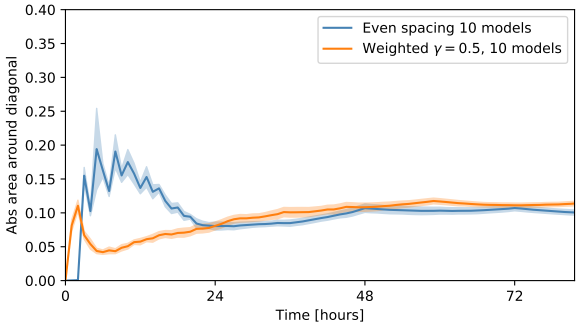

Figure 6: The area around the diagonal of the calibration plot for different variants of spacing of base models, evaluated individually for each horizon τ. Lower values correspond to better calibration. The error bands denote the standard error across the 5 splits.

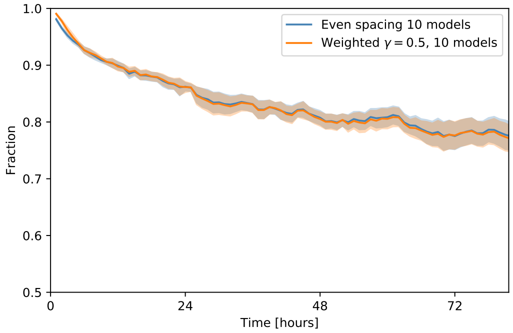

Figure 7: The fraction of concordant pairs per time τ: $|\{(i,j) \in P_\tau | S_i(\tau) < S_j(\tau)\}|/|P_\tau|$ for different spacing variants of base models. This gives an indication of discrimination performance as a function of the predictive horizon. Higher values represent better discrimination. The error bands denote the standard error across the 5 splits.