A study on intensive care early event prediction: How well do clinicians perform against AI?

Manuel Burger[1] **Jérôme Pasquier**[2] **Nadine C. Monai**[3] **Quinten Johnson**[1]

Xinrui Lyu[6] **Dinara Veshchezerova**[1] **Anastasia Escher**[2] **Carmen A. Pfortmueller**[3]

Joerg C. Schefold[3] **Tobias Merz**[4] **David Berger**[5] **Martin Faltys**[3]

Gunnar Rätsch[1*]

[1]Department of Computer Science, ETH Zurich, Switzerland

[2]NEXUS Personalized Health, ETH Zurich, Switzerland

[3]Department of Intensive Care Medicine, University Hospital and University of Bern, Switzerland

[4]Cardiovascular Intensive Care Unit, Auckland City Hospital, Auckland, New Zealand

[5]Department of Intensive Care Medicine, University Hospital Basel, University of Basel, Switzerland

[6] Swiss Data Science Center, ETH Zurich, Switzerland

Abstract

Machine learning models show strong capability in predicting impending organ failure by integrating high-volume dynamic patient data, but whether they can outperform clinicians is uncertain. Here, we compare the performance of ICU clinicians in predicting imminent circulatory and respiratory failure with previously published machine learning models. Our early results indicate that machine learning models deliver notable gains relative to clinicians and can meaningfully complement clinician judgment to potentially enhance patient care.

1 Introduction

Organ failure is a frequent and life-threatening complication in critically ill patients. Mortality is high: 30-50% for acute respiratory failure [1] and up to 50% for circulatory failure [2, 3]. Thus, these patients require an intensive care unit (ICU) stay, where continuous monitoring of organ function parameters allows early detection of deterioration and timely initiation of appropriate interventions [2].

ICU clinicians intermittently re-evaluate patients to determine whether any medical intervention is needed. In addition, ICUs are equipped with continuous monitoring systems that trigger alerts when predefined physiological thresholds are surpassed. While these threshold-based alarms are designed to draw clinicians' attention to potential deterioration, they often generate excessive number of alerts, many of which are false positive or clinically non-actionable, contributing to alarm fatigue. This desensitization of caregivers against alarms may hinder timely clinical-decision-making by obscuring

*Corresponding author: gunnar.raetsch@inf.ethz.ch

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

clinically relevant alarms [4, 5]. One promising approach is the integration of machine learning (ML) models to generate intelligent alarm algorithms [5]. For example, Hyland et al. [6] developed an early-warning system for circulatory failure, reducing alarm frequency up to 80 times compared to standard threshold-based systems.

ML-based alarm systems analyze continuously evolving patient data and integrate them in a prediction, triggering a warning before patients' deterioration [5]. Particularly in the data-rich environment of the ICU, such ML-models hold high potential for improving both patient care and resource management. Interest in ML-based predictive tools has grown rapidly in recent years, especially in the early detection of complications such as circulatory [6], respiratory [7, 8], renal failure [9–11], and sepsis [12], as well as mortality prediction [11] and forecasting the length of ICU stays [13]. Such models are predominantly trained and validated on large open-source retrospective datasets [8, 10, 14, 13]. However, their predictive performance often declines when applied to prospectively collected clinical data and new patient cohorts, raising concerns about their generalizability [10, 15]. Furthermore, only a few studies discuss the real-world implementation of these algorithms in clinical practice or assess their impact on patient outcomes [13, 16]. We are only aware of one study that directly compared the predictive accuracy of ML models against that of ICU clinicians in forecasting organ failure risk [17]. Additionally, few studies have looked into how ML predictions need to be presented to clinicians to best fit their needs [18].

This study aims to assess the ability of ICU clinicians to predict the risk of impending organ failure and to compare it to the performance of previously published ML-based risk prediction models. The findings of this study provide new evidence for the clinical utility of ML-based risk scores and support further research aimed at optimizing intelligent alarm systems for routine healthcare use.

2 Study Design and Methods

We conducted an observational single-center prospective cohort study between November 2024 and April 2025 in the interdisciplinary adult ICU of University Hospital Bern (Inselspital) in Switzerland with 393 admissions.

Study population: All adult patients ($\geq$18 years) admitted as emergency cases with invasive blood pressure monitoring were eligible. Exclusion criteria were refusal of secondary data use, primary neurologic admission, mechanical circulatory support or extracorporeal membrane oxygenation, and end-of-life care. For re-admissions, only the first ICU stay was included.

Objectives: The primary objective was to evaluate how accurately ICU physicians predict imminent circulatory failure within the 8-hour prediction horizon, compared with ML models. Circulatory failure [6] was defined as arterial lactate $\geq$2 mmol/L with either mean arterial pressure $\leq$65 mmHg or use of vasopressor/inotrope. Secondary objectives included prediction of respiratory failure (P/F ratio $\leq$200 mmHg) within the 24-hour prediction horizon [7].

Clinicians' assessments: Structured risk assessment questionnaires (Appendix B) were completed by the patients' treating physicians at randomized time points during the first 72 hours after admission to the ICU. The clinicians had the option of clinically reassessing the patient prior to filling out the questionnaires. The evaluations covered all shifts and weekdays. Participants included residents, attendings, and senior ICU physicians. Each prediction consisted of a binary estimate of the occurrence of organ failure, plus a self-reported confidence score. These were combined into a probability estimate (0–100%). Assessments had to be completed within one hour of notification.

Machine learning predictions: Each provided clinician assessment was paired with the prediction by a task-specific trained ML model, where model inputs are derived from routinely collected data such as vital monitors, laboratory tests and administered treatments extracted from the EPIC electronic health record (EHR) data management system. For circulatory failure, we reproduce the *CircEWS* (Gradient-Boosted-Tree-based [19]) model proposed by Hyland et al. [6]. For respiratory failure, we follow the modeling approach of Hüser et al. [7] who introduced *RMS* (Respiratory Monitoring System). In both cases, we rely on a "light" version of the model as proposed by the authors, which reduces the variable set considered for the model input and hence is anticipated to be more robust against distribution shifts [20].

We retrained both predictors on the HiRID-II [7] dataset, an update to the original HiRID [21] dataset, containing patient admissions from 2008 to 2019 with over 55,000 admissions. We use identical data

processing, annotation, and hyperparameter configurations as proposed by the published models and validate their performance on the HiRID-II dataset (Appendix Fig. 2). The study cohort was purely used for evaluation purposes. Given the temporal shift and the shift in the patient population (the study considered only emergency patient admissions while the training set incorporates intermediate care patients) between the training set and the study set, we observed a slight decrease in performance (see Appendix Fig. 2 A-B). The circulatory predictor is well calibrated and maintains this property also on the study evaluation set (Appendix Fig. 2 C). The respiratory predictor shows strong calibration on a hold-out test set of the training distribution (from the HiRID dataset) but shows slight overconfidence on the study dataset (Appendix Fig. 2 D).

Calibrating human risk assessments: Human expert annotations are expensive to obtain, especially in a high-stake environment such as intensive care. After collecting assessments from 64 participating physicians, we obtained a total of 3,145 time steps with assessments, which was filtered down to 2,327 when removing assessments performed during an annotated event, but only around 100 and 200 positively labeled early event prediction time steps for circulatory and respiratory failure, respectively (assessment was performed within the prediction horizon of the event). Hence, there were only few (in the tens) assessments provided by each clinician (see Appendix Fig. 3). The challenging part is that the clinicians are not calibrated against each other, each clinician has different confidence levels and sensitivity to observed signals. Pooling assessments from all clinicians might underestimate the true collective discriminative power of clinicians due to the misalignment of their confidence levels. We provide an overview of our current score alignment approach in Appendix C.

3 Results

Comparing ML and clinician risk assessments: The ML model, CircEWS, demonstrated substantially superior discriminative performance compared to treating physicians in predicting circulatory failure (Fig. 1 A). CircEWS achieved an area under the receiver operating characteristic curve (AuROC) of 0.843, while treating physicians achieved an AuROC of 0.634, representing over 20% improvement in discriminative ability ($p \ll 0.001$ using DeLong's test [22], we observe similar significance levels when accounting for clustering effects on patients or physicians [23]). For respiratory failure prediction, we observed a similar pattern albeit with a smaller performance differential (Fig. 1 C). The RMS model achieved an AuROC of 0.743 compared to 0.586 for treating physicians, representing over 15% improvement ($p \ll 0.001$ using DeLong's test).

Impact of Clinician Score Alignment: To address potential heterogeneity in clinician confidence calibration, we applied our monotonic alignment algorithm (Appendix C) to harmonize risk assessments across physicians. This alignment procedure significantly improved the collective predictive performance of clinicians (Fig. 1 D-E). For circulatory failure, the aligned physician scores achieved an AuROC of 0.706, representing an almost 7% improvement over the raw pooled assessments (0.634) and notably reduced the performance gap with the ML model but still being outperformed with a statistically significant margin ($p \ll 0.001$). Individual physician performance, when averaged across clinicians, was 0.650, suggesting that pooling without alignment underestimates collective clinical judgment. For respiratory failure, alignment showed modest gain, with aligned scores achieving an AuROC of 0.615 compared to 0.586 for raw assessments and 0.568 for mean individual physician performance.

Temporal Dynamics of Predictive Performance: We evaluated model and clinician performance across different prediction horizons to understand how temporal proximity to events affects predictive accuracy (Fig. 1 F-G). For circulatory failure at a matched precision of 0.10 (clinicians do not achieve considerably higher precision at any threshold, never above 0.20), the ML model achieved high recall further away from the event and increases to a 100% recall closer to the event. Meanwhile, clinicians showed high recall (expected at low precision) but with higher variance when bootstrapping across patients and unstable detection performance as we approach the event, suggesting no performance improvement with proximity to the event. For respiratory failure, the clinicians could achieve higher levels of precision and we matched ML model and clinicians at a fixed 0.50 precision. At this precision level, we saw a clear trend of improvements in recall for the ML model. The clinicians showed a slight increase in mean recall, but the variance observed across patient bootstraps suggests that this is not significant. We noticed that for patients with at least one failure event, clinicians could match the ML model in specificity at matched recall (sensitivity) for respiratory failure. This indicates that both the ML model and the clinician have a similar false alarm rate for critical patients

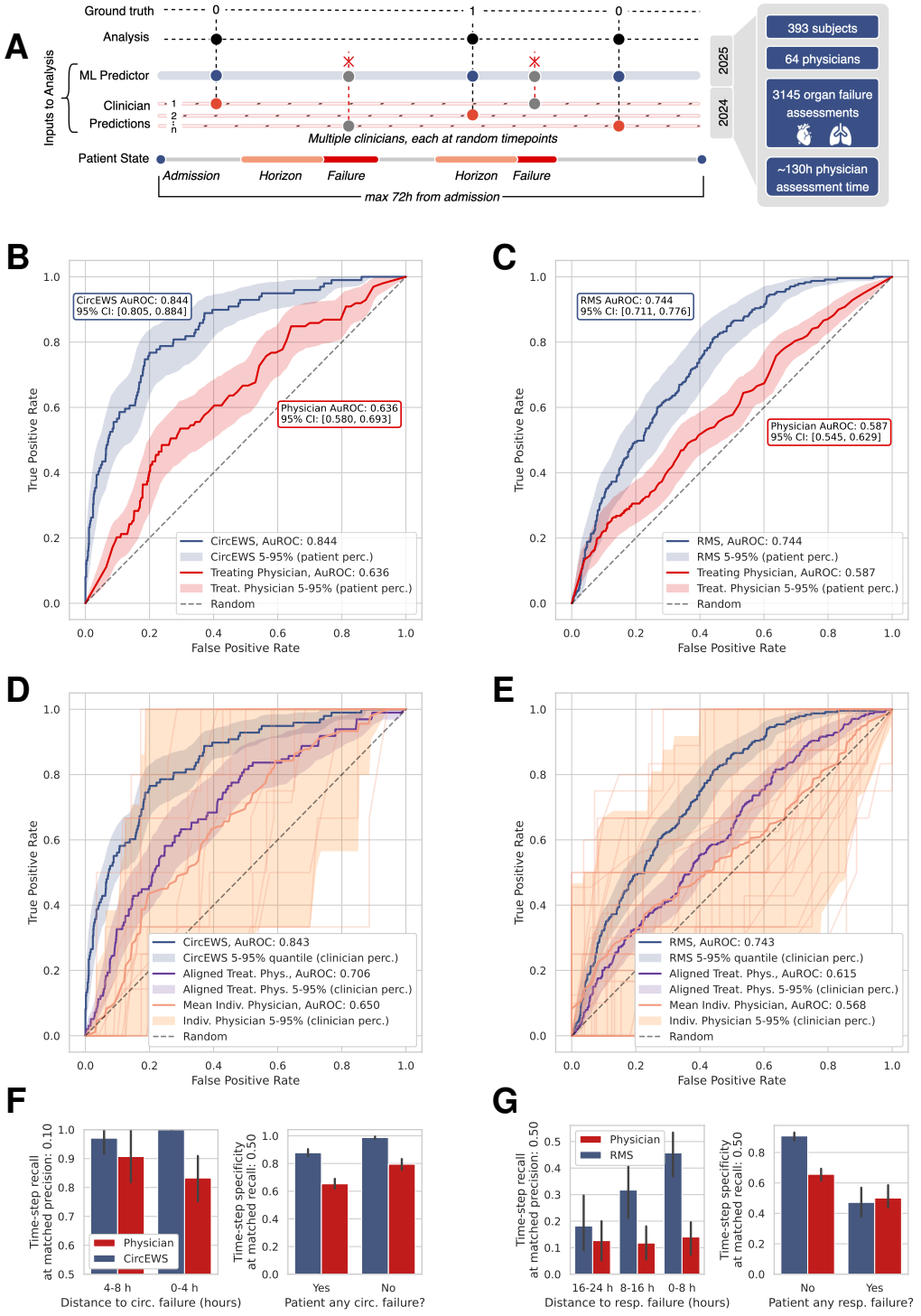

Figure 1: **A)** Study setup, comparing sparse clinician assessments against matched time point ML scores. Positive 1 labels are within the prediction horizon of a failure event and negative labels (stable) 0 are outside the prediction horizon of upcoming events. **B)** Circulatory failure assessments comparing CircEWS [6] against treating physicians. **C)** Respiratory failure assessments comparing RMS [7] against treating physicians. **D)** Circulatory failure assessments comparing CircEWS against individual physicians and aligned physician scores. **E)** Respiratory failure assessments comparing RMS against individual physicians and aligned physician scores. **F/G)** Circulatory/respiratory failure comparing binary assessments by physicians with ML predictions at matched precision over different prediction horizons (left plot); specificity at matched recall split by patients experiencing at least one or no failure event during their stay.

outside the event horizon, but the ML model better captures the signs of impending failure events and can still improve detection performance close to the event.

4 Discussion

Model Generalization to Study Cohort: Despite the temporal distribution shift between the training data (2008-2019) and the prospective study period (2024-2025), both ML models maintained robust performance (Appendix Fig. 2 A-B). CircEWS showed excellent calibration on the study cohort, while RMS exhibited slight overconfidence but retained strong discrimination (Appendix Fig. 2 C-D). These results support the generalizability of the ML models to recent ICU populations, albeit trained on historical data. However, we invested significant effort in meticulous data extraction, preprocessing, and alignment across the two datasets to ensure robust performance transfer. More robust and scalable models will facilitate the development of ML models for deployment [24].

Localizing risk: We hypothesize that clinicians tend to form baseline risk estimates rather than temporally localizing imminent risk, which may explain the relatively constant performance irrespective of proximity to the event.

Conclusion: The observed performance gaps have important clinical implications. While prior works [6, 9, 7] have established retrospectively that ML models can have strong early event prediction performance in intensive care, our preliminary study results suggest that they might meaningfully complement clinicians by raising awareness of acute deterioration in patients at risk of organ failure. We further anticipate to develop deeper insights on how ML models should provide predictions and what exactly they should predict to optimally complement the clinicians' existing discriminative capabilities.

References

[1] Giacomo Bellani, John G Laffey, Tài Pham, Eddy Fan, Laurent Brochard, Andres Esteban, and et al. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA*, 315(8):788–800, 2016.

[2] Maurizio Cecconi, Laura Evans, Mitchell Levy, and Andrew Rhodes. Sepsis and septic shock. *The Lancet*, 392(10141):75–87, 2018.

[3] Daniel D Berg, Erin A Bohula, and David A Morrow. Epidemiology and causes of cardiogenic shock. *Current Opinion in Critical Care*, 27(4):401–408, 2021.

[4] W Marc, HK Dirk, K Andreas, K Christian, S Wolfgang, and R Rainer. Alarm fatigue: Causes and effects. In *Studies in Health Technology and Informatics*, volume 243, pages 107–111. IOS Press, 2017.

[5] Martin Borowski, Matthias Görges, Robert Fried, Odette Such, Christian Wrede, and Michael Imhoff. Medical device alarms. *Biomedizinische Technik/Biomedical Engineering*, 56(2):73–83, 2011.

[6] Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3): 364–373, 2020.

[7] Matthias Hüser, Xinrui Lyu, Martin Faltys, Alizée Pace, Marine Hoche, Stephanie Hyland, Hugo Yèche, Manuel Burger, Tobias M Merz, and Gunnar Rätsch. A comprehensive ml-based respiratory monitoring system for physiological monitoring & resource planning in the icu. *medRxiv*, 2024. doi: 10.1101/2024.01.23.24301516. URL `https://www.medrxiv.org/content/early/2024/01/23/2024.01.23.24301516`.

[8] Son T Le, Emilie Pellegrini, Abigail Green-Saxena, Courtney Summers, Jared Hoffman, Jacob Calvert, and et al. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ards). *Journal of Critical Care*, 60:96–102, 2020.

[9] Xinrui Lyu, Bowen Fan, Matthias Hüser, Philip Hartout, Thomas Gumbsch, Martin Faltys, Tobias M Merz, Gunnar Rätsch, and Karsten Borgwardt. An empirical study on kdigo-defined acute kidney injury prediction in the intensive care unit. *Bioinformatics*, 40(Supplement_1): i247–i256, 2024.

[10] Irene Vagliano, Nicholas C Chesnaye, JH Leopold, Kitty J Jager, Ameen Abu-Hanna, and Marcel C Schut. Machine learning models for predicting acute kidney injury: a systematic review and critical appraisal. *Clinical Kidney Journal*, 15(12):2266–2280, 2022.

[11] M Syed, S Syed, K Sexton, HB Syeda, M Garza, M Zozus, and et al. Application of machine learning in intensive care unit (icu) settings using mimic dataset: Systematic review. *Informatics*, 8(1):16, 2021.

[12] Michael Moor, Nicolas Bennett, Drago Plečko, Max Horn, Bastian Rieck, Nicolai Meinshausen, Peter Bühlmann, and Karsten Borgwardt. Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *EClinicalMedicine*, 62, 2023.

[13] David Shillan, Jonathan AC Sterne, Alan Champneys, and Benedict Gibbison. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Critical Care*, 23(1):284, 2019.

[14] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, and et al. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1): 160035, 2016.

[15] Patrick Rockenschaub, Andreas Hilbert, T Kossen, P Elbers, F von Dincklage, VI Madai, and et al. The impact of multi-institution datasets on the generalizability of machine learning prediction models in the icu. *Critical Care Medicine*, 52(11):1710–1721, 2024.

[16] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340, 2019. doi: 10.1038/s41591-019-0548-6. URL https://doi.org/10.1038/s41591-019-0548-6.

[17] Marina Flechet, Stefano Falini, Carlo Bonetti, Fernando Güiza, Miet Schetz, Greet Van den Berghe, and et al. Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the akipredictor. *Critical Care*, 23(1):282, 2019.

[18] Melissa D McCradden, Kelly Thai, Azadeh Assadi, Sana Tonekaboni, Ian Stedman, Shalmali Joshi, Minfan Zhang, Fanny Chevalier, and Anna Goldenberg. What makes a 'good' decision with artificial intelligence? a grounded theory study in paediatric care. *BMJ Evidence-Based Medicine*, 30(3):183–193, 2025. ISSN 2515-446X. doi: 10.1136/bmjebm-2024-112919. URL https://ebm.bmj.com/content/30/3/183.

[19] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[20] Malte Londschien, Manuel Burger, Gunnar Rätsch, and Peter Bühlmann. Domain generalization and adaptation in intensive care with anchor regression, 2025. URL https://arxiv.org/abs/2507.21783.

[21] M. Faltys, M. Zimmermann, X. Lyu, M. Hüser, S. Hyland, G. Rätsch, and T. Merz. HiRID, a high time-resolution icu dataset (version 1.1.1). *PhysioNet*, 2021. doi: 10.13026/nkwc-js72. URL https://doi.org/10.13026/nkwc-js72.

[22] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, sep 1988.

[23] Nancy A. Obuchowski. Nonparametric Analysis of Clustered ROC Curve Data. 53(2):567–578. ISSN 0006-341X. doi: 10.2307/2533958. URL https://www.jstor.org/stable/2533958.

[24] Manuel Burger, Daphné Chopard, Malte Londschien, Fedor Sergeev, Hugo Yèche, Rita Kuznetsova, Martin Faltys, Eike Gerdes, Polina Leshetkina, Peter Bühlmann, and Gunnar Rätsch. A foundation model for intensive care: Unlocking generalization across tasks and domains at scale. *medRxiv*, 2025. doi: 10.1101/2025.07.25.25331635. URL https://www.medrxiv.org/content/early/2025/07/25/2025.07.25.25331635.

[25] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999.

[26] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, Bonn, Germany, 2005. ACM.

[27] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, 1964. doi: 10.1093/comjnl/7.2.155.

A Ethics

Ethical approval was obtained from the Institutional Review Board of the Canton of Bern (BASEC project ID: 2024-01046).

B Risk Assessment Questionnaire for Clinicians

No.	Question [Options for answers]
1	Treating or non-treating clinician? [treating, non-treating]
2	Was an additional clinical assessment of the patient performed, and if so, how was it conducted?' [bedside, Epic (EHR) only, no additional assessment]
3	When was the patient last clinically assessed? [now, 15-30 min, 30-60 min, >1 hour]
4	Will the patient experience circulatory failure according to the study definition in the next 8 hours? [yes, no]
5	How confident are you in your above assessment regarding circulatory failure? [0-100]
6	Will the patient experience respiratory failure according to the study definition in the next 24 hours? [yes, no]
7	How confident are you in your above assessment regarding respiratory failure? [0-100]
8	What is the estimated survival time of this patient since admission to the intensive care unit? [<28 days, 1-6 months, 6-12 months, >1 year]
9	What were the two most important clinical variables considered in your prediction of circulatory failure?

B.1 Clinician scores

Clinicians provided binary predictions (yes/no) with confidence scores (0-100%). These were converted to probabilities: $P(failure) = 0.5 \pm 0.5 \times (confidence/100)$, where the sign depends on the binary prediction. For example, "yes" with 80% confidence yields 0.9, while "no" with 60% confidence yields 0.2. The alignment procedure (Appendix C) further harmonizes scores across clinicians.

C Calibration and Score Alignment

To fix the underestimation of the discriminative power of a collective of clinicians, the risk assessments provided by clinicians must be aligned. This could be achieved by treating each clinician like a separate prediction model and applying common calibration techniques. However, we quickly discovered that due to the small set of labels and especially very scarce number of positive labels (many clinicians even without an assessment of a positive label) common calibration methods such as Platt-scaling [25] or isotonic regression [26] are impractical to apply. We are actively researching this challenging risk score alignment problem and share our current framework for aligning the scores of different clinicians. To avoid using (scarce) labels for calibration, we propose to use the machine learning risk scores as a shared target space for aligning clinicians (note that we show the model itself is well-calibrated). This avoids degenerated solutions for clinicians with very few to no positively labeled assessments and provides a calibrated shared space to align the scores from different clinicians. We use Powell's optimization algorithm [27] to fit a monotonically increasing set of scores for each clinician, minimizing the mean absolute error against the machine learning model's predicted probabilities.

C.1 Mathematical Formulation

Let $\mathcal{C} = \{1, \ldots, K\}$ denote the set of K clinicians. For each clinician $c \in \mathcal{C}$, we observe:

- A set of annotations $\mathbf{a}^{(c)} = \{a_1^{(c)}, \ldots, a_{n_c}^{(c)}\}$ where $a_i^{(c)} \in [0, 1]$ represents the risk assessment for patient i

- Corresponding model predictions $\mathbf{m}^{(c)} = \{m_1^{(c)}, \ldots, m_{n_c}^{(c)}\}$ where $m_i^{(c)} \in [0, 1]$

Our goal is to find a monotonic transformation function $f_c : [0, 1] \to [0, 1]$ for each clinician c that aligns their annotations to the model's probability space. Due to the discrete nature of clinician annotations, we parameterize f_c by learning transformed values for each unique annotation level for a given clinician c.

Let $\mathbf{u}^{(c)} = \{u_1^{(c)}, \ldots, u_{L_c}^{(c)}\}$ be the sorted unique annotation values for the clinician c, where $u_1^{(c)} < u_2^{(c)} < \ldots < u_{L_c}^{(c)}$. We seek to learn transformed values $\tilde{\mathbf{u}}^{(c)} = \{\tilde{u}_1^{(c)}, \ldots, \tilde{u}_{L_c}^{(c)}\}$ such that:

$$\tilde{\mathbf{u}}^{(c)*} = \underset{\tilde{\mathbf{u}}^{(c)} \in [0,1]^{L_c}}{\arg\min} \ \text{MAE}(\tilde{\mathbf{u}}^{(c)}, \mathbf{m}^{(c)}) \quad \text{subject to:} \tag{1}$$

$$\tilde{u}_1^{(c)} \leq \tilde{u}_2^{(c)} \leq \ldots \leq \tilde{u}_{L_c}^{(c)} \quad \text{(monotonicity)} \tag{2}$$

$$0 \leq \tilde{u}_j^{(c)} \leq 1 \quad \forall j \in \{1, \ldots, L_c\} \quad \text{(bounded)} \tag{3}$$

where the mean absolute error is computed as:

$$\text{MAE}(\tilde{\mathbf{u}}^{(c)}, \mathbf{m}^{(c)}) = \frac{1}{n_c} \sum_{i=1}^{n_c} \left| \tilde{u}_{I(a_i^{(c)})}^{(c)} - m_i^{(c)} \right| \tag{4}$$

and $I(a_i^{(c)})$ maps the annotation $a_i^{(c)}$ to its corresponding index in $\mathbf{u}^{(c)}$.

C.2 Algorithm Overview

Algorithm 1 Monotonic Alignment of Clinician Annotations

Require: Dataset $\mathcal{D}$ with annotations, model scores, and clinician IDs
Ensure: Aligned annotations for all clinicians

Initialize $\mathcal{D}_{aligned} \leftarrow \mathcal{D}$ with empty aligned annotation column
$\mathcal{C} \leftarrow$ unique clinician IDs in $\mathcal{D}$
for each clinician $c \in \mathcal{C}$ **do**
  $\mathcal{D}_c \leftarrow$ subset of $\mathcal{D}$ for clinician c
  $\mathbf{a}^{(c)} \leftarrow$ annotations from $\mathcal{D}_c$
  $\mathbf{m}^{(c)} \leftarrow$ model scores from $\mathcal{D}_c$
  Sort pairs $(\mathbf{a}^{(c)}, \mathbf{m}^{(c)})$ by annotations
  $\mathbf{u}^{(c)} \leftarrow$ unique sorted values from $\mathbf{a}^{(c)}$
  Create index mapping: $I_i \leftarrow$ index of $a_i^{(c)}$ in $\mathbf{u}^{(c)}$
  Initialize: $\tilde{\mathbf{u}}_{init}^{(c)} \leftarrow \mathbf{u}^{(c)}$ ▷ Use original values as initial guess
  Define objective:

$$J(\tilde{\mathbf{u}}^{(c)}) = \begin{cases} \frac{1}{n_c} \sum_{i=1}^{n_c} |\tilde{u}_{I_i}^{(c)} - m_i^{(c)}| & \text{if constraints satisfied} \\ \infty & \text{otherwise} \end{cases}$$

  Optimize: $\tilde{\mathbf{u}}^{(c)*} \leftarrow$ Powell(J, $\tilde{\mathbf{u}}_{init}^{(c)}$)
  for each sample i in $\mathcal{D}_c$ **do**
    $\mathcal{D}_{aligned}[i]$.aligned_annotation $\leftarrow \tilde{u}_{I_i}^{(c)*}$
  end for
end for
return $\mathcal{D}_{aligned}$

C.3 Implementation Details

The optimization problem is solved using Powell's conjugate direction method [27], which is derivative-free and suitable for our constrained optimization setting. The monotonicity constraint is enforced within the objective function by returning an infinite cost when violated. This approach ensures that:

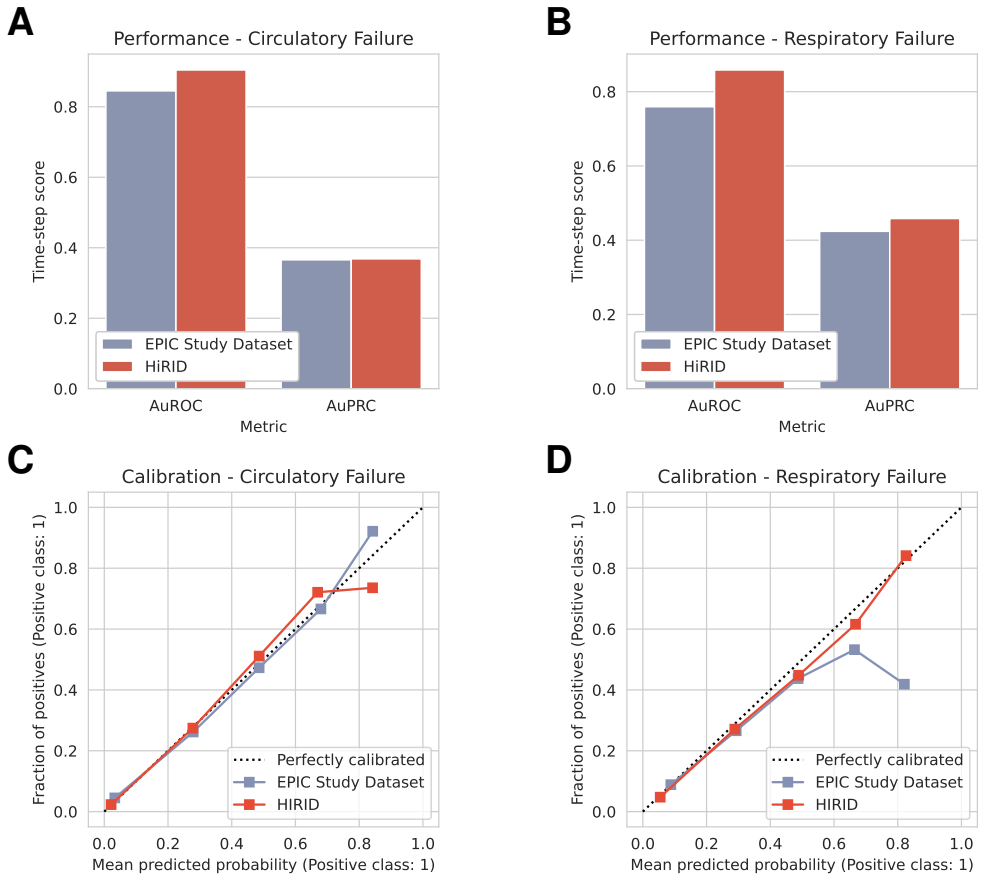

Figure 2: **A/B)** Model performance for circulatory/respiratory failure on the test set of the training source (HiRID-II [7]) and the study evaluation dataset extracted from the EPIC electronic health record system in use at the study center, which was not used for training. **C/D)** Model calibration of the circulatory/respiratory failure predictor on the training source (HiRID-II [7]) and the study evaluation dataset extracted from the EPIC electronic health record system.

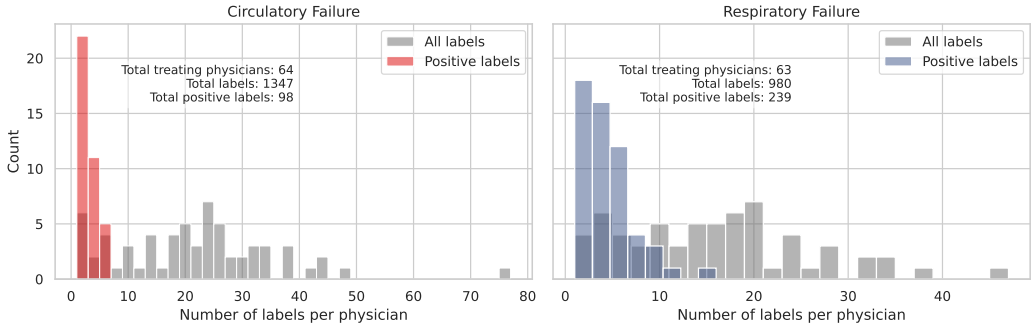

Figure 3: **Assessment Distribution** Number of assessments (labels) per participating treating physician during periods which were included in the final analysis (e.g. outside of failure events)

1. Each clinician's ranking of patients is preserved (monotonicity)
2. The transformed scores are aligned with the well-calibrated model predictions
3. No ground truth labels are required for the alignment process
4. Clinicians with few assessments can still be effectively aligned

The mean absolute error (MAE) is chosen over mean squared error (MSE) as it is more robust to outliers.