Yet Another ICU Benchmark (YAIB)

Robin van de Water[1], Hendrik Schmidt[1], Paul Elbers[2], Patrick Thoral[2], Bert Arnrich[1], Patrick Rockenschaub[3]

[1]Hasso Plattner Institute, University of Potsdam [2]Amsterdam UMC, Vrije Universiteit [3]Lab for AI in Medicine, Charité — Universitätsmedizin Berlin

ICLR International Conference On Learning Representations

ICU Data: complex, sparse, error-prone

- ICU data is a large collection of **sparse and irregularly sampled events** (e.g., lab tests).
- Information is frequently encoded using local, **non-standard terminologies**.
- Medical concepts of interest (e.g., sepsis) are not directly recorded but need to be retrospectively derived from these events.

⚠ ICU data is often processed and analyzed in different, non-reproducible ways. ⚠

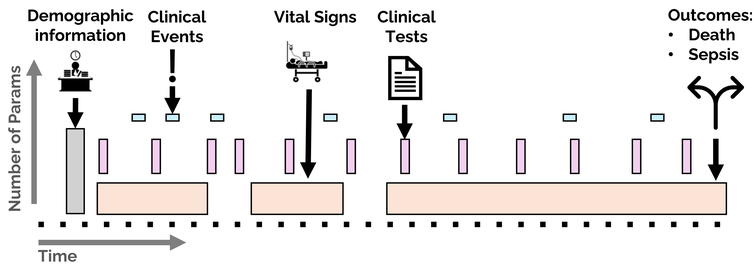

Figure 1. *Schematic timeline overview of ICU data for a single patient.*

Why *yet another* ICU benchmark?

- Despite efforts for standardization, there remain **substantial variations in definitions** of features and cohorts among studies.
- Existing ICU benchmarks are limited to 1-2 datasets with hard-coded tasks that provide **little support for extensibility and cross-dataset comparisons** (1; 2).

☛ *YAIB* provides a modular framework for multi-dataset ICU prediction.

Table 1. *Supplemental details of openly accessible ICU datasets.* Accessing each dataset requires completing a credentialing procedure, although smaller demo versions are available for MIMIC and eICU (number of demo stays shown in parentheses).

Dataset	MIMIC-III / IV	eICU CRD	HiRID	AUMCdb	*Your dataset*
Stays	40k (0.1k)/ 73k	201k (2k)	34k	23k	
Frequency (time)	1 hour	5 minutes	2 / 5 minutes	up to 1 minute	
Origin	USA	USA	Switzerland	Netherlands	
Published	2015	2017	2020	2019	
Benchmarks	3	1	1	0	
Tasks	Classification: Mortality, Acute Kidney Injury, Sepsis Regression: Kidney Function, Length of Stay				*Your tasks*

Design philosophy: reproducibility *and* extensibility first

Guiding principles:

- Medical research is inherently **complex**.
- One size fits all is **impossible**.
- Hardcoded solutions are **not reproducible**.

Desiderata:

- **Out-of-the-box support** for datasets, tasks, models enabling quick prototyping.
- Reproducible **modular** setup to change only what you need (*Configuration as Code*).
- Full extensibility **across experiment lifecycle**: dataset, cohorts, preproc, model, metrics.

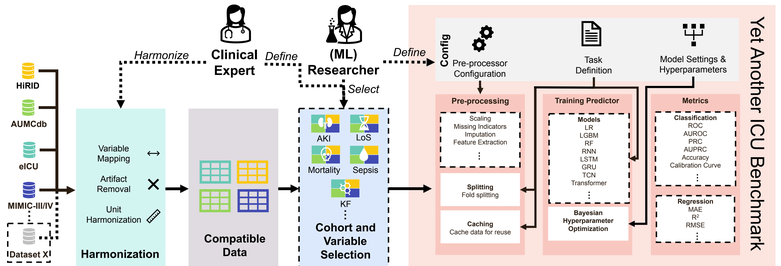

Figure 2. *Schematic overview of benchmark pipeline.* On the left side, the creation of harmonized ICU cohorts is shown. Domain expertise of clinicians is crucial to define clinically useful tasks. The schematic overview of the benchmark stages can be found on the right. Note that the dotted line indicates that this component can be easily extended, as it follows an abstracted interface.

Cohort definitions can heavily skew prediction accuracy

- Sepsis is a **clinically relevant** endpoint where early treatment could prevent death.
- There are **several different sepsis definitions** based on clinical values (3).

☛ Differences in cohort definitions dramatically impact model performance.

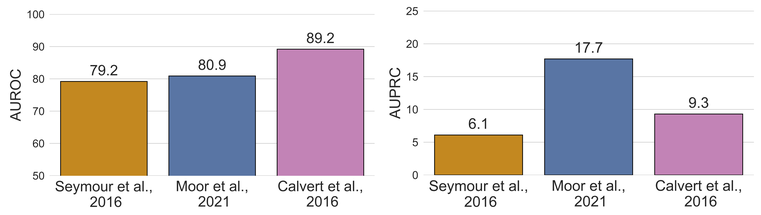

Figure 3. *Sepsis prediction on MIMIC-IV for different definitions of sepsis.* **Left:** AUROC, **Right:** AUPRC.

Important steps for clinical implementation are now easy as pie

- **External validation** — a crucial step in establishing the reliability of a model — is rarely done because of a lack of multi-center data (4).
- **Fine-tuning** can provide a solution to data shortages when developing a local model.

☛ In YAIB, external validation and fine-tuning are readily available.

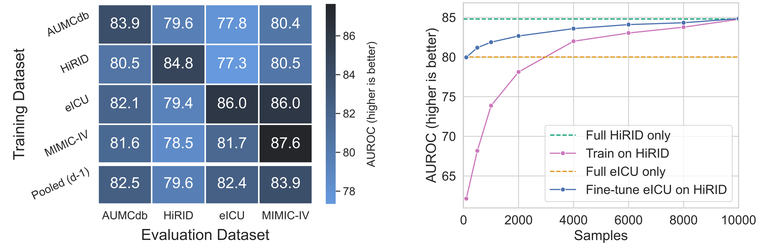

Figure 4. **Left**: Performance in AUROC of GRU models predicting ICU mortality when trained on one dataset (rows) and evaluated on all others (columns). Pooled (d-1) refers to training a model on every dataset except the evaluation dataset. **Right**: Fine-tuning of a GRU model for ICU mortality trained on eICU for prediction on HiRID.

Takeaways

- Status Quo: studies have their own ad-hoc pipelines with a 1) specific dataset, 2) cohort definition, and 3) preprocessing pipeline; $\Longrightarrow$ **each choice has a major impact on the results (often more than a SOTA model).**
- *YAIB* aids researchers by providing them with **ready-to-use datasets, endpoints, and models**; new models can therefore be easily **compared and validated**.
- While most existing benchmarking studies are hard-coded, we utilize flexible, **dataset-independent** cohort definitions and configurable preprocessing facilities linked via a **common, shareable syntax.**

YAIB
Yet Another **ICU Benchmark**

Paper Github

References

[1] Chaoqi Yang, Zhenbang Wu, Patrick Jiang et al. PyHealth: A Deep Learning Toolkit for Healthcare Applications. KDD '23. Association for Computing Machinery, August 2023.
[2] Hugo Yèche, Rita Kuznetsova, Marc Zimmermann et al. HiRID-ICU-Benchmark — A Comprehensive Machine Learning Benchmark on High-resolution ICU Data. NeurIPS '22.
[3] Michael Moor, Bastian Rieck, Max Horn et al. Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. *Frontiers in Medicine*, 8, 2021.
[4] Patrick Rockenschaub, Ela Marie Akay, Benjamin Gregory Carlisle et al. Generalisability of ai-based scoring systems in the icu: a systematic review and meta-analysis, 2023.